Ostbayerische Technische Hochschule Amberg-Weiden
Fakultät Elektrotechnik, Medien und Informatik

Studiengang Künstliche Intelligenz

Masterarbeit

von

Tobias **Schotter**

Vorhersage von Jobkündigungen anhand von Lohnabrechnungsdaten

Prediction of job resignations based on payroll data

Ostbayerische Technische Hochschule Amberg-Weiden
Fakultät Elektrotechnik, Medien und Informatik

Studiengang Künstliche Intelligenz

Masterarbeit

von

Tobias **Schotter**

Vorhersage von Jobkündigungen anhand von Lohnabrechnungsdaten

Prediction of job resignations based on payroll data

Bearbeitungszeitraum:	von	2. Mai 2023
	bis	30. Oktober 2023

1. Prüfer:	Prof. Dr.-Ing Christoph P. Neumann
2. Prüfer:	Prof. Dr. Fabian Brunner

Ostbayerische Technische Hochschule Amberg-Weiden
Fakultät Elektrotechnik, Medien und Informatik

Ostbayerische Technische Hochschule
Amberg-Weiden

Selbstständigkeitserklärung

Name und Vorname
der Studentin/des Studenten: **Schotter, Tobias**

Studiengang: **Künstliche Intelligenz**

Ich bestätige, dass ich die Masterarbeit mit dem Titel:

Vorhersage von Jobkündigungen anhand von Lohnabrechnungsdaten

selbständig verfasst, noch nicht anderweitig für Prüfungszwecke vorgelegt, keine anderen als die angegebenen Quellen oder Hilfsmittel benutzt sowie wörtliche und sinngemäße Zitate als solche gekennzeichnet habe.

Datum: 30. November 2023

Unterschrift:

Ostbayerische Technische Hochschule Amberg-Weiden
Fakultät Elektrotechnik, Medien und Informatik

Masterarbeit Zusammenfassung

Studentin/Student (Name, Vorname):	**Schotter, Tobias**
Studiengang:	Künstliche Intelligenz
Aufgabensteller, Professor:	Prof. Dr.-Ing Christoph P. Neumann
Durchgeführt in (Firma/Behörde/Hochschule):	DATEV eG
Betreuer in Firma/Behörde:	Dr. Frank Eichinger
Ausgabedatum: 2. Mai 2023	Abgabedatum: 30. Oktober 2023

Titel:

Vorhersage von Jobkündigungen anhand von Lohnabrechnungsdaten

---

Zusammenfassung:

Diese Masterarbeit untersucht die Möglichkeit, Arbeitnehmerkündigungen mithilfe von Lohnabrechnungsdaten vorherzusagen. Ein Hauptaugenmerk liegt auf der Frage der Umsetzbarkeit und dem entscheidenden Thema des Feature Engineerings. Dabei werden Daten in Millionenhöhe verarbeitet. Im Rahmen dieser Untersuchung erfolgt ein Vergleich zwischen Random Forests und neuronalen Netzen. Zur Gewährleistung einer zuverlässigen Evaluation wird neben den gängigen maschinellen Lernmetriken auch ein naives Modell entwickelt, das einzig auf dem Unterschied zwischen dem Jahresgehalt und der Marktwertprognose basiert. Die Arbeit untersucht, ob die entwickelten Modelle geeignet sind, um im Bereich des proaktiven Personalmanagements eingesetzt zu werden. Die frühzeitige Identifizierung von Mitarbeitern, die möglicherweise das Unternehmen verlassen werden, ermöglicht gezielte Maßnahmen zur Stärkung ihrer Bindung an das Unternehmen.

Die Ergebnisse zeigen, dass es zwar nicht möglich ist, jede Kündigung präzise im Voraus vorherzusagen, aber in vielen Fällen bereits gute Vorhersagen möglich sind. Dabei hat sich herausgestellt, dass das neuronale Netzwerk bessere Ergebnisse in Bezug auf verschiedene Metriken liefert. Infolgedessen wird ein mögliches Anwendungsszenario für ein solches Vorhersagemodell in einer Beratungsanwendung für Steuerberater erörtert. Um Steuerberatern eine bessere Grundlage für das Verständnis dieser Vorhersagen zu bieten, wird das Thema Explainable AI angesprochen, und es wird eine Technik vorgestellt, um Kündigungsvorhersagen besser nachvollziehbar zu gestalten.

Inhaltsverzeichnis

Kapitel 1

Hintergrund und Motivation

Die Bindung qualifizierter Mitarbeiter ist für Unternehmen von zentraler Bedeutung. Eine hohe Fluktuation in einem Unternehmen kann zu erheblichen Kosten und Produktivitätsverlusten führen. Der Verlust wertvoller Fachkräfte kann nicht nur zu einem Mangel an Know-how und Erfahrung führen, sondern auch zu einem Rückgang der Unternehmensleistung und der Servicequalität. Angesichts der zunehmenden Konkurrenz um Talente und der steigenden Anforderungen an eine dynamische Arbeitswelt müssen Unternehmen Strategien entwickeln, um ihre Mitarbeiter langfristig an das Unternehmen zu binden. In diesem Zusammenhang soll nun die Vorhersage von Arbeitnehmerkündigungen mithilfe von Datenanalyse und maschinellem Lernen als Hilfestellung leisten. Indem Vorhersagemodelle implementiert werden könnten, die frühzeitig potenzielle Kündigungen identifizieren, eröffnen sich wertvolle Chancen zur proaktiven Gestaltung der Mitarbeiterbindung. Diese gezielten Interventionsstrategien tragen dazu bei, das Vertrauen und die Loyalität der Mitarbeiter zu stärken und eine positive Unternehmenskultur zu etablieren. Die Schaffung einer produktiveren und zufriedeneren Arbeitsumgebung durch die Verwendung von Datenanalyse und maschinellem Lernen geht somit über die reine Vorhersage von Kündigungen hinaus. Indem Unternehmen ihre Mitarbeiter besser verstehen und ihre Bedürfnisse antizipieren, können sie innovative Ansätze zur Mitarbeiterbindung entwickeln und ihre Wettbewerbsfähigkeit auf dem Arbeitsmarkt stärken.

1.1 Problemstellung und Zielsetzung

In dieser Masterarbeit wird untersucht, inwieweit die Lohnabrechnungsdaten der DATEV eG genutzt werden können, um Arbeitnehmerkündigungen vorherzusagen. DATEV eG ist der drittgrößte Anbieter für Business-Software in Deutschland (IDC-Ranking 2023) [54] und einer der großen europäischen IT-Dienstleister, wobei sich der Datenbestand der Lohnabrechnungsdaten auf über 14.5 Millionen Einträgen pro Monat summiert. Darüber hinaus zeichnet sich dieser mit einem besonders verlässlichen Datenursprung (Echtdaten aus Lohnbuchhaltung) aus. Auf diesem Datenbestand sollen nun verschiedene Methoden der Datenanalyse und des maschinellen Lernens

angewandt werden, um die bestmögliche Vorhersagegenauigkeit für dieses Szenario zu ermitteln. Darüber hinaus werden die ethischen Implikationen der Verwendung von Mitarbeiterdaten diskutiert und mögliche Lösungen zur Wahrung der Privatsphäre der Mitarbeiter vorgestellt. Das Ergebnis dieser Arbeit wird dazu beitragen, Unternehmen bei der Entwicklung von Strategien zur Mitarbeiterbindung zu unterstützen und somit einen Beitrag zur Schaffung einer produktiveren und zufriedeneren Arbeitsumgebung zu leisten.

Obwohl es bereits Studien zur Vorhersage von Arbeitnehmerkündigungen gibt, unter anderem [55, 44, 4, 53], bleibt die Frage offen, wie effektiv diese Methoden wirklich sind. Denn die dort verwendeten Datensätze sind in der Regel sehr klein oder synthetisch generiert, wodurch eine Anwendung auf zukünftige Echtdaten keine Resultate garantiert. Nun soll diese Lücke in der Forschung untersucht werden, wie gut ein maschinelles Lernmodell die Kündigung von Mitarbeitern anhand von produktiven Lohnabrechnungsdaten vorhersagen kann. Hier besteht die Möglichkeit auf einen Datenbestand von kleinen, mittleren und großen Unternehmen in Millionengröße zuzugreifen. Es ist daher von Interesse zu untersuchen, welche Merkmale für die Vorhersage von Arbeitnehmerkündigungen am relevantesten sind (Feature Selection) und wie ein maschinelles Lernmodell auf der Grundlage dieser Daten entwickelt werden kann, um genaue Vorhersagen zu liefern. Im Rahmen der Untersuchung wird auch eine umfassende Evaluation durchgeführt, um festzustellen, ob das entwickelte Modell in der Lage ist, eine Vorhersage für einen Zeitraum von 12 Monaten in die Zukunft zu liefern. Es ist wichtig anzumerken, dass dies keine triviale Aufgabe ist und nicht erwartet werden kann, dass jede Kündigung korrekt vorhergesagt wird. Dennoch können bereits teilweise korrekte Vorhersagen von Kündigungen wertvolle Erkenntnisse liefern. Ähnliche Herausforderungen sind auch in der Forschung zu Churn Predictions (Vorhersage von Kundenabwanderungen) zu beobachten [1, 2, 6], da dort ähnliche Ansätze verwendet werden. Selbst in diesen Forschungsbereichen ist es schwierig, perfekte Vorhersagen zu erzielen, da es sich um eine grundsätzlich komplexe Aufgabe handelt.

Zunächst sollen relevante Variablen identifiziert bzw. konstruiert werden, die eine Vorhersage von Kündigungen ermöglichen. Hierbei sollen insbesondere Lohnabrechnungsdaten, sowie weitere relevante Faktoren, welche aus diesen gewonnen werden können, darunter beispielsweise die Beschäftigungsdauer, Distanz zum Arbeitgeber oder die Position im Unternehmen berücksichtigt werden. Auf Basis dieser identifizierten Variablen sollen Vorhersagemodelle entwickelt werden, die eine möglichst hohe Genauigkeit (Performance) bei der Vorhersage von Kündigungen erreichen. Hierbei wird untersucht, welcher potenzielle Nutzen aus den Ergebnissen erwachsen kann und wie das entwickelte Modell möglicherweise in ein bereits existierendes Produkt integriert werden kann. Durch diese Analyse sollen sowohl die Leistungsfähigkeit des Modells als auch seine praktische Anwendbarkeit und mögliche Mehrwerte für bestehende Lösungen beleuchtet werden. Da die Verwendung von Lohnabrechnungsdaten auch ethische Implikationen hat, wird eine Diskussion über den Datenschutz und die Wahrung der Privatsphäre der Mitarbeiter geführt. Mögliche Lösungen werden vorgestellt, um sicherzustellen, dass die Verwendung von Daten ethisch vertretbar ist. Durch

die Erfüllung dieser Ziele soll die Masterarbeit einen Beitrag zur Verbesserung der Mitarbeiterbindung leisten und Unternehmen dabei helfen, ihre wertvollen Mitarbeiter langfristig an sich zu binden. Hierbei möchte ich meine bisherige Erfahrung im Bereich der Softwarekonstruktion darlegen, die ich während vorherigen Projekten gesammelt habe [7, 50, 28].

In den kommenden Kapiteln dieser Arbeit werden wir uns mit einer umfangreichen Untersuchung zu dem Thema Mitarbeiterkündigungen und Gehaltsvorhersage befassen. In Kapitel 2 werden verwandte Arbeiten und differenziert vorgestellt. Kapitel 3 gibt einen Überblick über die methodischen Grundlagen des maschinellen Lernens, einschließlich relevanter Begriffserklärungen und der Auswahl geeigneter Lernmodelle. Kapitel 4 behandelt die Beschaffung und Vorbereitung der Daten, einschließlich der Methoden zur Datenerhebung, -bereinigung und -analyse, während Kapitel 5 die methodische Vorgehensweise und die verwendeten Analysestrategien erläutert. In Kapitel 6 werden die Ergebnisse dieser Untersuchung im Zusammenhang mit den Forschungsfragen diskutiert. Abschließend fasst Kapitel 7 die wichtigsten Erkenntnisse zusammen und geht auf das zukünftige Vorgehen ein, inklusive eines abschließendem Fazits.

Kapitel 2

Verwandte Arbeiten

Der Stand der Forschung zu maschinellen Lernmodellen in verschiedenen Bereichen von Kündigungsvorhersagen ist umfassend untersucht wurden. Hierbei liegt der Schwerpunkt allerdings auf der Vohersage von Kundenkündigungen. Zwar gibt es eine Reihe von Studien die sich mit der Identifizierung von Faktoren und Mustern beschäftigen, welche mit Mitarbeiterkündigungen in Verbindung stehen, jedoch liegt der Schwerpunkt meist nur auf einzelnen Teilaspekten. Sehr häufig wird der Begriff der *Churn Predictions* (auf Deutsch auch Kündigungsvorhersage oder Kundenabwanderungsvorhersage genannt) verwendet. Dazu bezieht sich der Begriff *Churners* auf Kunden, die ein Unternehmen verlassen oder abwandern. Im Geschäftsumfeld ist dies ein gängiger Terminus, der Kunden kennzeichnet, die ihre Bindung an ein Unternehmen verlieren, sei es durch Kündigung eines Abonnements, das Beenden einer Mitgliedschaft oder den Wechsel zu einem Konkurrenten. Grundsätzlich handelt es sich um ein Verfahren bei dem Datenanalysen und statische Modelle verwendet werden, um vorherzusagen, welche Kunden ein Unternehmen verlassen werden. Also Schwerpunkt ist es Kunden oder Abonnenten zu erkennen, die sich entscheiden, ein Produkt oder eine Dienstleistung eines Unternehmens nicht mehr zu nutzen. Die *Churn Predictions*-Modelle nutzen unter anderem historische Daten über Kundenverhalten, Transaktionen, sowie demografische Informationen, welche grundlegend von den für diese Arbeit zur Verfügung stehenden Daten abweichen. Dennoch sind einige dazu aufgelisteten Strategien von Relevanz, wodurch eine Auflistung dieser Arbeiten von Bedeutung ist.

2.1 Vorherige Studien zu Churn Predictions

Eine Studie im Bereich von Churnpredictions in der Telekommunikationsbranche [29] hat gezeigt, wie maschinelles Lernen erfolgreich eingesetzt werden kann, um Kündigungen vorherzusagen. Obwohl sich diese Studie nur auf die Telekommunikationsbranche konzentriert, sind die zugrunde liegenden Konzepte und Methoden auch auf die allgemeine Vorhersage von Mitarbeiterkündigungen übertragbar. Die Autoren verwendeten verschiedene maschinelle Lernalgorithmen, darunter Entschei-

dungsbäume, neuronale Netzwerke und Support Vector Machines, um Churn-Muster anhand von Kundendaten zu identifizieren. Die Ergebnisse zeigten, dass maschinelles Lernen ein effektives Mittel sein kann, um Kündigungen vorherzusagen und Strategien zur Kundenbindung zu entwickeln. Hier muss allerdings bedacht werden, dass einige verwendete Features nicht auf Lohnabrechnungsdaten übertragbar sind, wie Beschwerdeinformationen der Kunden, welche einen großen Einfluss auf das Modell haben können. Der hierbei verwendete Datensatz umfasste knapp 400.000 Einträge, was vergleichsweise zu anderen Studien in diesem Feld sehr viel ist, allerdings nur ein Bruchteil der Daten, welche in dieser Arbeit zur Verfügung stehen. Eine weitere Studie im Bereich der Telekommunikation [34] hat eine viel versprechende Architektur gezeigt, allerdings wurde das Verhältnisse der Klassen zwischen Churners und nicht Churners nicht erläutert, wodurch die Qualität der Ergebnisse nicht zu 100% eingestuft werden kann.

Eine weitere viel beachtete Studie im Bereich Churn Predictions der Telekommunikationsbranche von Huang et al. [30], zeigte ebenfalls einen Ansatz für die Vorhersage von Kundenabwanderungen an einem Produktivdatensatz im Millionen-Bereich eines Mobilfunkbetreibers in China. Sie erzielten eine Genauigkeit (Precision) von 0.96 für die ersten 50.000 vorausgesagten Churners. Darunter haben sie eine außergewöhnliche Technik gegen Datenunausgewogenheit (Weighted Instance) angewandt. Sie ordnen jeder Instanz ein proportionales Gewicht zu, wobei Churners ein höheres Gewicht und Nicht-Churners ein niedrigeres Gewicht zugewiesen bekommen. Durch diese Technik haben sich die Modelle um etwa 10% gegenüber der nicht ausgeglichenen Methode verbessert. Es hat sich gezeigt, dass Random Forest etwas besser als andere Klassifikatoren abschneidet. Ebenso stellen sie fest, dass die Wahl des Klassifikators eher zweitrangig ist, weil die Merkmale (Features) einen deutlich höheren Einfluss auf das Resultat aufweisen. Bei einer Vielzahl von Merkmalen können die meisten skalierbaren Klassifikatoren fast die gleiche Genauigkeit (accuracy) erreichen. Das Ergebnis zeigt, dass das Hinzufügen eines guten Merkmals die prädiktive Modellierung deutlicher verbessern kann als der Wechsel zu einem anderen Klassifikator.

Adbelrahim et al. [2] hat verschiedene baumbasierte Algorithmen für die Vorhersage von Kundenkündigungen verglichen, dabei unter anderem Entscheidungsbäume, Random Forest, GBM-Baumalgorithmus und XGBoost. In der vergleichenden Analyse schnitt XGBoost besser ab als die anderen in Bezug auf die AUC-Genauigkeit [11]. Praveen et al. [6] haben eine vergleichende Analyse von Modellen bereitgestellt, in welcher Algorithmen zu Support Vector Machines, Entscheidungsbäumen, Naive Bayes und logitische Regression für *Churn Predictions* verglichen wurden. Anschließend analysierten sie auch die Wirkung von Boosting-Algorithmen auf die Klassifizierungsgenauigkeit. Die Ergebnisse zeigen, dass SVM-POLY mit AdaBoost die besten Leistungen erzielen konnte. Horia Beleiu et al. [12] hat drei Ansätze des maschinellen Lernens verglichen darunter Neuronales Netze, Support Vector Machines und Bayes'sche Netze für kundenbasierte *Churn Predictions*. Bei der Auswahl der Trainings-Features wurde Principle Component Analysis (PCA) [37] verwendet, um die Dimensionen der Daten zu reduzieren. Für die Leistungsbewertung wurde ebenso die AUC-Genauigkeit verwendet [11]. J. Burez et al. [16] haben versucht, das Problem des Klassenungleichgewichts zu

erfassen. Sie verwendeten logistische Regression und Random Forest mit Re-Sampling Techniken. Es hat sich herausgestellt, dass das Problem des Klassenungleichgewichts durch den Einsatz von optimierungsbasiertem Stichprobenverfahren reduziert werden kann.

Das Forschungsgebiet im Bereich der Customer Churn Predictions ist vielfältig. Obwohl es sich insbesondere in Bezug auf die Datenstruktur von Employee Churn Predictions unterscheidet, konnten dennoch wertvolle Erkenntnisse gewonnen werden. Besonders baumbasierte Klassifikatoren haben sich als leistungsstarke Ansätze in diesem Problemfeld erwiesen. Es wurde auch gezeigt, dass die Wahl des Klassifikators im Vergleich zur Auswahl der Features von sekundärer Bedeutung ist. Es ist jedoch wichtig zu erwähnen, dass die Ergebnisse der Studie von Huang et al. [30] genau zu betrachten sind: Ihre Verwendung von Features wie „long distance minutes in the past 3 months“ ist in Europa aus Datenschutzgründen nicht erlaubt und könnte somit die Vergleichbarkeit der Ergebnisse beeinträchtigen.

2.2 Vorherige Studien zu Employee Churn Predictions

Neben einer Reihe von Studien im Bereich von *Customer Churn Predictions*, wurde auch das Szenario der *Employee Churn Predictions* von verschiedenen Forschern untersucht [55, 44, 4, 53]. Yue Zhao et al. [55] befasste sich mit der Untersuchung verschiedener Klassifikationsalgorithmen und bewertete die Fähigkeit dieser zur Vorhersage von Mitarbeiterfluktuation. Ebenfalls betrachteten sie Probleme die oftmals von anderen Studien in diesem Forschungsgebiet außer Acht gelassen werden sehr kritisch. Einer der Aspekte, die hervorgehoben wurde, ist die traditionelle Verwendung von Accuracy-Metriken als führende Evaluationsmaßnahme. Jedoch erweisen sich diese als unzuverlässig, wenn es um unausgewogene Datensätze geht. [3, 48, 49]. Dies resultiert aus der Tatsache, dass in der Regel der Anteil der Personen, die eine Organisation verlassen, kleiner ist als derjenige, der bleibt. Dadurch entsteht die Gefahr, dass Ergebnisse mit einer irreführend hohen Accuracy berechnet werden können. Ein weiteres aufgeführtes Problem dieser Studie ist der oftmalige Versuch die Interpretierbarkeit des Modells durch die Einstufung der Feature Importance (Relevanz der Merkmale) zu verbessern. Die Analyse der Feature Importance in mehreren Studien [5] könnte voreingenommen sein, da sie klassifikatorabhängige Ansätze verwendet bei denen die Modellleistung mit einfließt. Ein Beispiel dafür besteht darin, dass Feature Importance durch Entscheidungsbäume mithilfe einer internen eingebundenen Funktion, beispielsweise der Implementierung von *scikit-learn*, berechnet werden könnten. Wenn die Entscheidungsbäume jedoch keine guten Leistungen erbringen, kann das entsprechende Ergebnis ungenau sein. Eine Alternative besteht in der Berechnung der Feature Importance durch Permutation. Allerdings ist diese Methode sehr rechenaufwändig, insbesondere bei großen Datensätzen mit vielen Merkmalen.

Punnoose, R. et al [44] hat verschiedene Algorithmen für die Vorhersage von Mitarbeiterkündigungen verglichen und kam zu dem Ergebnis, dass der XGBoost Klassifikator [17] ein überlegener Algorithmus im Bezug auf Genauigkeit, Laufzeit und Speichernutzung im Rahmen von Churn Prediction ist. Zusätzlich zeichnet sich dieser Klassifikator

durch eine höhere Robustheit hinsichtlich der Handhabung von Datenrauschen im Vergleich zu anderen Klassifikatoren aufgrund seiner Regularisierung aus. Alamsyah A. et al [4] hat ebenfalls verschiedene Klassifikatoren verglichen, darunter Naive Bayes, Decision Tree und Random Forest anhand des Human Resources Datensatzes [43]. Hierbei hat sich der Random Forest Klassifikator als der genauste herausgestellt. Wobei in dieser Studie erneut nicht auf das Verhältnis zwischen Churners und nicht Churners eingegangen ist, wodurch eine Accuracy von 97,5% nicht viel Aussagekraft hat. Zusätzlich dazu verfügt dieser Datensatz lediglich über knapp 16.000 Datenpunkte, was darauf hindeutet, dass das Modell vermutlich nicht effektiv auf noch unbekannte Daten angewendet werden kann. Yigit I. O. et al. [53] hat verschiedene maschinelle Lernmethoden für den Fall der Employee Churn Predictions an dem IBM HR Analytics Employee Attrition & Performance Datensatz [43] verglichen. Dennoch weist dieser Datensatz eine hohe Generizität auf und enthält nur wenige Einträge von Merkmalen, die in einem realen Szenario schwer zugänglich sind, wie beispielsweise *Relationship Satisfaction*, *Work Life Balance* und *Performance Ratings*. Es ist außerdem zu beachten, dass der Datensatz synthetisch generiert wurde

Trotz der oben erwähnten Bandbreite an Forschungsergebnisse sind die Erkenntnisse von Methoden des maschinellen Lernens zur Vorhersage von Mitarbeiterfluktuation oft problemspezifisch und schwer zu verallgemeinern. Dies liegt in erster Linie daran, weil Personaldaten vertraulich sind, was naturgemäß die Durchführung eingehender Analysen über mehrere Datensätze erschwert. Darüber hinaus sind Personaldaten oft verrauscht, inkonsistent und enthalten fehlende Informationen [18, 44]. Ein Problem, das durch den geringen Anteil der Mitarbeiterfluktuation, der in der Regel in einem bestimmten Datensatz von Personaldaten vorhanden ist, noch verschärft wird.

Der große Unterschied dieser Arbeit zu den oben aufgelisteten Forschungsbeiträgen ist der Fakt, dass sich der hier verwendete Datensatz nicht nur auf ein einzelnes Unternehmen oder einen spezifischen fiktiven Fall bezieht, sondern auf eine Bandbreite von kleinen, mittleren und großen Unternehmen in ganz Deutschland.

2.3 Vorhersage von Gehältern

Neben einer Reihe von Studien im Bereich der Vorhersage von Mitarbeiterkündigungen, ist die Studie von Eichinger F. et al. [23] im Rahmen von Gehaltsvorhersagen ebenso erwähnenswerte. In dieser Studie wurde untersucht, ob und weshalb ein fortschrittlicher maschineller Lernansatz – nämlich Random Forest Regression – qualitativ hochwertige Gehaltsvorhersagen erreichen kann. Dafür wurde ein realer Datensatz aus dem Bereich der Lohnabrechungen verwendet. Der Ansatz lernt für jeden Beruf ein Random Forest Regressionsmodell zur Vorhersage der Gehälter. Es wurde gezeigt, dass dieser Ansatz besser als verwandte Arbeiten zur Gehaltsvorhersage abschneidet.

Diese Studie verdient Beachtung, da sie auf einer sehr ähnlichen Datenquelle basierte wie diese Masterarbeit. Beide Studien nutzen Gehaltsabrechnungen von etwa drei bis vier Millionen Arbeitnehmern pro Monat über einen Zeitraum von einem Jahr. Diese Daten stammen aus der Lohnabrechnungssoftware der deutschen Firma DATEV eG

und umfassen mehr als 300 verschiedene Berufe. Ein erfolgreiches Ergebnis an der selben Datenbasis liefert wertvolle Einblicke in das betrachtete Forschungsgebiet. Hier ist noch anzumerken, dass die Methoden, welche in dieser Studie beschrieben wurden nochmals überarbeitet wurden und das Modell noch besser als beschrieben im Produktivsystem der Abteilung PBo (DATEV Personal-Benchmark online) eingesetzt wird. Genauer gesagt wird dort nun ein neuronales Netz aufbauend auf drei verschiedenen Modellen verwendet. Darunter ein kategoriales, numerisches und boolesches Modell, welche miteinander verkettet werden.

Ein weiteres Produkt für die Vorhersage von Gehältern ist *Gehaltsvergleich BETA*, von dem Statistischen Bundestamt [15] bereitgestellt. Es ermöglicht Gehaltsvergleiche für verschiedene Berufe und Branchen durchzuführen. Es basiert auf umfangreichen Daten aus Lohnabrechnungen und bietet eine Informationsquelle für die Analyse von Gehaltsstrukturen in Deutschland. Die Lohnabrechnungen stammen von Arbeitgebern in Deutschland und werden monatlich an die Sozialversicherung gemeldet. Dabei werden Gehaltsinformationen von Millionen von Beschäftigten erfasst, was eine breite und repräsentative Datengrundlage darstellt. Die Daten umfassen Gehälter, Berufe, Branchen und Arbeitszeiten. Gehaltsvergleich BETA bietet verschiedene Funktionen, die es den Nutzern ermöglichen Gehaltsvergleiche durchzuführen. Nutzer können Gehälter nach verschiedenen Kriterien filtern, wie zum Beispiel nach Berufserfahrung, Bildungsniveau oder Unternehmensgröße. Zudem ermöglicht das Tool die Darstellung von Gehaltsentwicklungen über einen bestimmten Zeitraum hinweg. Insgesamt liefert der Gehaltsvergleich BETA von DESTATIS wertvolle Erkenntnisse über Gehaltsstrukturen in Deutschland und trägt zur Verbesserung der Transparenz auf dem Arbeitsmarkt bei. Das Kapitel 4 wird detailliertere Informationen über den genutzten Datensatz liefern, einschließlich der Methoden zur Datenerhebung, -bereinigung und -analyse.

Kapitel 3

Methodische Grundlagen - Machine Learning

In dieser Arbeit werden verschiedene überwachte Algorithmen des maschinellen Lernens (Supervised Learning) beschrieben, demonstriert und hinsichtlich ihrer Fähigkeit zur Vorhersage der Mitarbeiterfluktuation bewertet. Dieses Kapitel gibt einen allgemeinen Überblick über die Theorie hinter diesen Algorithmen und den Grundkonzepten nötig für diese.

3.1 Untergruppen der künstlichen Intelligenz

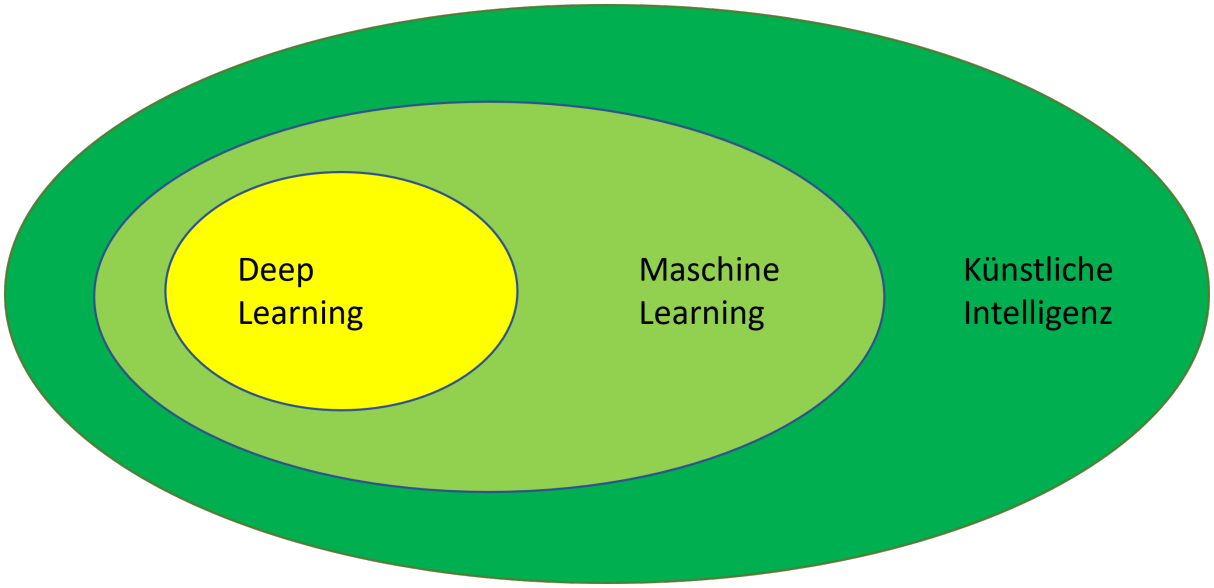


Abbildung 3.1: Deep Learning, Machine Learning und KI

Im allgemeinen Sprachgebrauch werden die Begriffe Künstliche Intelligenz, Machine Learning und Deep Learning oft als Synonyme verwendet. Es ist jedoch wichtig zu

betonen, dass es sich nicht um identische Begriffe handeln, sondern eher um verschiedene Untergruppen innerhalb des übergeordneten Bereichs der künstlichen Intelligenz. Vergleiche Abbildung 3.1 für eine visualle Darstellung dieser Untergruppen.

3.1.1 Künstliche Intelligenz (Artificial Intelligence)

Künstliche Intelligenz (KI) ist ein multidisziplinäres Forschungsgebiet, das sich mit der Entwicklung von Systemen befasst, die in der Lage sind, menschenähnliches Denken und Verhalten zu demonstrieren. Das Hauptziel der KI besteht darin, intelligente Maschinen zu schaffen, die in der Lage sind, Aufgaben auszuführen, die normalerweise menschlicher Intelligenz bedürfen. Dies umfasst Fähigkeiten wie Lernen, Problemlösung, Mustererkennung, Sprachverarbeitung, Wahrnehmung und Entscheidungsfindung [45, 41, 38].

Die Begründung für das Forschungsgebiet der KI geht auf die Studie von McCarthy, Minsky, Rochester und Shannon (1955) zurück, die als "A Proposal for the Dartmouth Summer Research Project on Artificial Intelligence" bekannt ist. In dieser Arbeit wurde der Grundstein für die Entwicklung der KI als eigenständige Disziplin gelegt und die Vision formuliert, intelligente Maschinen zu schaffen [38].

In der KI werden verschiedene Techniken und Ansätze verwendet, darunter symbolische Logik, Expertensysteme, neuronale Netze, genetische Algorithmen und maschinelles Lernen. Der Fokus liegt darauf, Algorithmen und Modelle zu entwickeln, die es Computern ermöglichen, Wissen zu erwerben, zu verarbeiten, zu interpretieren und in Bezug auf verschiedene Aufgaben und Anwendungen zu nutzen [45]. Künstliche Intelligenz findet Anwendung in vielen Bereichen, einschließlich der Medizin, Robotik, Automatisierung, Sprachverarbeitung, Bilderkennung, Spielentwicklung und Datenanalyse [45]. Die kontinuierliche Weiterentwicklung von KI-Technologien und die steigende Verfügbarkeit großer Datenmengen haben zu bedeutenden Fortschritten in diesem Feld geführt.
Obwohl Künstliche Intelligenz beeindruckende Fortschritte erzielt hat, bleiben jedoch einige Herausforderungen bestehen. Dazu gehören ethische Fragen im Zusammenhang mit Autonomie und Verantwortlichkeit von KI-Systemen, Datenschutz und Sicherheitsbedenken sowie die Angemessenheit und Transparenz der Entscheidungsfindung von KI-Systemen. Insgesamt spielt die Künstliche Intelligenz eine wichtige Rolle bei der Gestaltung und Weiterentwicklung unserer modernen Technologien [45, 41, 38].

3.1.2 Maschinelles Lernen (Machine Learning)

Maschinelles Lernen (ML) ist ein Teilgebiet der Künstlichen Intelligenz, das sich mit der Entwicklung von Algorithmen und Modellen befasst, die es Computern ermöglichen, aus Erfahrungen zu lernen und automatisch Muster und Zusammenhänge in Daten zu erkennen [10, 27]. Im Gegensatz zu traditionellen programmierten Algorithmen, bei denen explizite Anweisungen zur Lösung eines Problems gegeben werden, basiert maschinelles Lernen auf der Idee, dass Computer aus Daten lernen können, um Vorhersagen oder Entscheidungen zu treffen [10]. Dieser Ansatz ermöglicht es, komplexe

Probleme zu bewältigen und Einsichten aus großen Datenmengen zu gewinnen.
Es gibt verschiedene Arten von ML-Algorithmen, darunter überwachtes Lernen (Supervised Learning), unüberwachtes Lernen (Unsupervised Learning) und bestärkendes Lernen (Reinforcement Learning). Beim überwachten Lernen werden Modelle mithilfe von Eingabe-Daten und den entsprechenden Ausgabe-Label (Zielvariablen) trainiert, um Vorhersagen zu treffen [27]. Beim unüberwachten Lernen werden hingegen Muster und Strukturen in den Daten entdeckt, ohne vorab bekannte Ausgabe-Label zu verwenden. Bestärkendes Lernen bezieht sich auf das Lernen von Entscheidungsstrategien basierend auf Rückmeldungen und Belohnungen [10, 39].
Maschinelles Lernen findet in einer Vielzahl von Anwendungsbereichen Anwendung, einschließlich Bilderkennung, Sprachverarbeitung, Textanalyse, medizinische Diagnose, Finanzprognosen und Empfehlungssysteme [10, 27]. Die hohe Flexibilität und Anpassungsfähigkeit von ML-Modellen ermöglicht es, komplexe Aufgaben zu automatisieren und genaue Vorhersagen zu treffen.

Trotz der großen Fortschritte im Bereich des maschinellen Lernens gibt es auch Herausforderungen zu beachten. Dazu gehören die Notwendigkeit qualitativ hochwertiger und umfangreicher Trainingsdaten, die Vermeidung von Überanpassung (Overfitting) der Modelle an die Trainingsdaten, die Interpretierbarkeit von ML-Modellen sowie ethische Fragen im Zusammenhang mit Datenschutz und Bias [27]. Insgesamt spielt maschinelles Lernen eine entscheidende Rolle bei der Datenanalyse und Automatisierung von Entscheidungsprozessen. Durch die Entwicklung fortschrittlicher ML-Modelle und -Algorithmen streben Forscher danach, neue Erkenntnisse aus Daten zu gewinnen und komplexe Probleme effektiv zu lösen [10, 27, 39].

3.1.3 Tiefes Lernen (Deep Learning)

Tiefes Lernen (Deep Learning) ist die letzte Untergruppe der künstlichen Intelligenz, welche auf künstlichen neuronalen Netzwerken basiert, die in der Lage sind, automatisch komplexe Muster und Merkmale aus großen Datenmengen zu extrahieren [25, 35].
Im Gegensatz zu traditionellen ML-Modellen, die auf manuell entwickelten Merkmalen basieren, sind tiefe neuronale Netzwerke in der Lage, hierarchische Merkmale direkt aus den Rohdaten zu extrahieren und komplexe Muster zu erlernen [25]. Dies wird durch die Verwendung von mehreren Schichten von Neuronen erreicht, die miteinander verbunden sind und die Fähigkeit haben, automatisch Merkmale in den Daten zu entdecken und zu generalisieren. Tiefes Lernen hat in den letzten Jahren beeindruckende Fortschritte erzielt und ist heute in vielen Anwendungsbereichen weit verbreitet. Es findet Anwendung in der Bilderkennung, Sprachverarbeitung, Sprachübersetzung, automatischen Fahrzeugsteuerung und vielen anderen Bereichen [25, 35]. Die Verwendung von tiefem Lernen hat zu erstaunlichen Ergebnissen geführt, wie beispielsweise die Fähigkeit von Computern, Bilder zu erkennen oder menschenähnliche Sprache zu generieren.

Es gibt jedoch auch Herausforderungen beim tiefen Lernen. Tiefe neuronale Netzwerke erfordern eine große Menge an Trainingsdaten und eine erhebliche Rechenleistung,

um effektiv zu funktionieren. Das Training von tiefen Modellen kann zeitaufwendig sein und erfordert spezialisierte Hardware, wie Grafikprozessoren (GPUs) oder Tensor Processing Units (TPUs). Zudem besteht die Gefahr der Überanpassung (Overfitting), bei dem das Modell zu stark auf die Trainingsdaten spezialisiert ist und nicht gut auf neue Daten generalisiert [25].
Trotz dieser Herausforderungen hat das tiefe Lernen bahnbrechende Ergebnisse in verschiedenen Bereichen erzielt und eröffnet neue Möglichkeiten für die Verarbeitung und Analyse großer Datenmengen. Die Forschung in diesem Bereich ist weiterhin aktiv, um die Leistungsfähigkeit und Effizienz von tiefen neuronalen Netzwerken zu verbessern [25, 35].

Nachdem nun die wesentlichen Begrifflichkeiten der Domäne „Künstliche Intelligenz“ erläutert und voneinander abgegrenzt wurden, folgt nun eine Erläuterung von Themen, die in deren Kontext adressiert werden müssen und im weiteren Verlauf der Arbeit eine Rolle spielen.

3.2 Concept Drift

In diesem Abschnitt wird das Phänomen des Concept Drifts erläutert, welches eine Herausforderung für das maschinelle Lernen, vor allem bei der Vorhersage von Geschehnissen in der Zukunft, darstellt. Relevanz hierfür besteht dadurch, dass dieses Problem in einem gewissen Rahmen innerhalb dieser Arbeit auftreten wird. Concept Drift tritt auf, wenn die statistischen Eigenschaften der Daten über die Zeit hinweg veränderlich sind und somit die Modelle und Algorithmen, die auf diesen Daten basieren, beeinflusst werden.

Ursachen und Arten von Concept Drift

Concept Drift bezieht sich auf die Veränderung der zugrunde liegenden Konzepte oder Verteilungen der Daten über die Zeit. Dies kann verschiedene Ursachen haben, wie z.B. Veränderungen in den Umweltbedingungen, Änderungen im Verhalten des Systems oder kontinuierliche Evolution von Mustern in den Daten. „Concept Drift stellt eine Herausforderung dar, da die Annahmen, die bei der Modellbildung gemacht werden, nicht mehr gültig sein können“ [24].

Es gibt verschiedene Arten von Concept Drift, die sich hinsichtlich der Art der Veränderungen unterscheiden. Ein häufiges Unterscheidungsmerkmal ist die plötzliche (sudden) versus die schleichende (gradual) Veränderung. Plötzlicher Concept Drift tritt auf, wenn die Änderungen abrupt erfolgen, während schleichender Concept Drift eine allmähliche Veränderung über die Zeit hinweg darstellt. Die Unterscheidung zwischen plötzlichem und schleichendem Concept Drift ist wichtig, da unterschiedliche Ansätze erforderlich sind, um mit diesen beiden Arten umzugehen [24].

Abbildung 3.2 verdeutlicht, dass ein Concept Drift nicht nur zu einem genauen Zeitpunkt stattfinden kann, sondern auch über einen Zeitraum andauern. Infolgedessen können während der Transformation Zwischenkonzepte auftreten, wenn ein Aus-

gangskonzept in ein anderes Endkonzept übergeht. Ein Zwischenkonzept kann eine Mischung aus Anfangs- und Endkonzept sein, wie bei einem inkrementellen (Incremental) Drift [36]. Es ist jedoch auch möglich, dass Konzepte wiederkehrend sind, wie das Reoccurring Concept. Ein Beispiel hierfür könnten Saisonalitäten sein, also Ereignisse, die nur in bestimmten Jahreszeiten innerhalb der Daten auftreten.

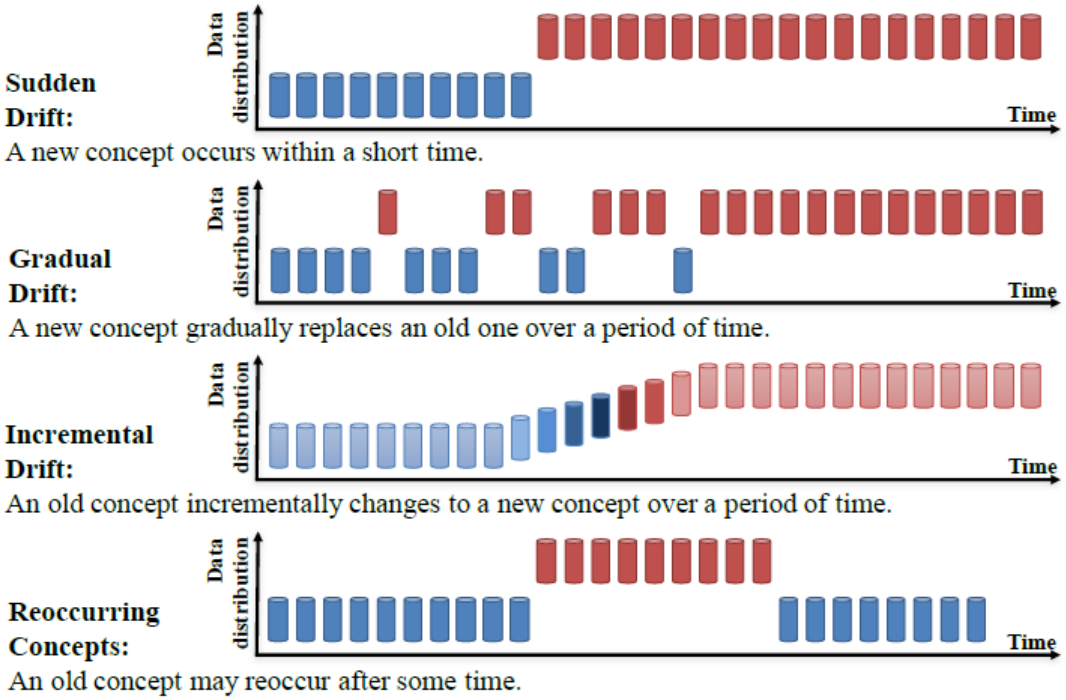


Abbildung 3.2: Ein Beispiel verschiedener Concept Drift Typen [36]

.

Auswirkungen und Techniken um Concept Drift zu vermeiden

Concept Drift stellt eine Herausforderung für das maschinelle Lernen dar, da die Modelle, die auf historischen Daten trainiert wurden, möglicherweise nicht mehr für die aktuellen Daten geeignet sind. Dies kann zu schlechter Vorhersageleistung und einer Abnahme der Modellgenauigkeit führen. „Concept Drift kann zu einem schleichenden Leistungsverlust von maschinellen Lernmodellen führen und erfordert daher die Entwicklung von Techniken zur Erkennung und Anpassung an Veränderungen“ [20].

Es gibt verschiedene Techniken, um mit Concept Drift umzugehen:

- **Überwachtes Neulernen:** Beim überwachten Neulernen wird das Modell periodisch mit aktuellen Daten erneut trainiert, um den Wissensstand des Modells auf dem neuesten Stand zu halten. Es werden neue Daten verwendet, um das Modell anzupassen und Veränderungen in den Daten zu berücksichtigen [24]. Dies ermöglicht es dem Modell, sich an sich ändernde Bedingungen anzupassen und Concept Drift zu erkennen.

- **Inkrementelles Lernen:** Beim inkrementellen Lernen wird das Modell schrittweise aktualisiert, indem neue Beispiele schrittweise hinzugefügt werden. Anstatt das Modell vollständig neu zu trainieren, werden neue Daten verwendet, um das Modell anzupassen und seine Fähigkeit zur Bewältigung von Concept Drift zu verbessern. Dieser Ansatz ermöglicht eine kontinuierliche Anpassung des Modells an sich ändernde Bedingungen [19].
- **Ensemble-Methoden:** Ensemble-Methoden kombinieren mehrere Modelle zu einem Gesamtmodell, um die Robustheit gegenüber Concept Drift zu erhöhen. Indem verschiedene Modelle miteinander kombiniert werden, können Schwächen einzelner Modelle ausgeglichen und die Vorhersagegenauigkeit verbessert werden. Ensemble-Methoden wie Bagging, Boosting und Stacking können verwendet werden, um die Leistung des Modells zu verbessern und die Auswirkungen von Concept Drift zu reduzieren [33].
- **Adaptive Lernrate:** Die adaptive Lernrate ist eine Technik, bei der die Lernrate des Modells basierend auf der Veränderung der Datenverteilung angepasst wird. Wenn Concept Drift erkannt wird, kann die Lernrate erhöht oder verringert werden, um die Gewichtungsaktualisierung entsprechend zu steuern. Dies ermöglicht es dem Modell, sich schneller an neue Daten anzupassen und Veränderungen in den Daten effektiver zu berücksichtigen [32].
- **Rekalibrierung:** Die Rekalibrierung bezieht sich auf die Anpassung der Klassenverteilung oder der Vorhersageschwelle des Modells, um Veränderungen in den Daten zu berücksichtigen. Wenn sich die Verteilung der Klassen oder die Bedingungen im Laufe der Zeit ändern, kann die Rekalibrierung helfen, die Vorhersagen des Modells anzupassen und die Genauigkeit zu verbessern. Diese Technik kann besonders nützlich sein, um die Leistung des Modells bei ungleich verteilten Klassen oder bei Veränderungen in den Kostenfunktionen zu optimieren [13].
- **Transfer Learning:** Transfer Learning bezieht sich auf die Verwendung von Wissen aus vorherigen Aufgaben, um die Anpassungsfähigkeit des Modells an neue Daten zu verbessern. Durch die Nutzung bereits erlernter Merkmale und Muster aus verwandten Aufgaben kann das Modell schneller und effektiver auf neue Aufgaben reagieren und Concept Drift minimieren. Transfer Learning ist besonders nützlich, wenn die neuen Daten nur begrenzt sind und ein umfangreiches Training auf neuen Daten nicht möglich ist [42].
- **Fensterbasierte Ansätze:** Fensterbasierte Ansätze verwenden nur die neuesten Daten zur Modellaktualisierung und verwerfen ältere Daten. Anstatt das gesamte historische Datenmaterial zu verwenden, werden nur die Daten in einem bestimmten Zeitfenster betrachtet. Dadurch kann das Modell schnell auf Veränderungen reagieren und sich an die aktuellen Bedingungen anpassen. Diese Methode ist besonders nützlich, wenn sich die Datenverteilung im Laufe der Zeit ändert [9].
- **Konzept-Drift-Erkennung:** Die Konzept-Drift-Erkennung umfasst die Überwa-

chung statistischer Maßnahmen, um das Auftreten von Concept Drift zu erkennen. Dazu gehören Maßnahmen wie die Überwachung der Genauigkeitsänderung, der Distanz zwischen aktuellen und vorherigen Modellen oder der Veränderung von Merkmalsverteilungen. Wenn Anzeichen für Concept Drift erkannt werden, kann das Modell entsprechend angepasst werden, um Veränderungen in den Daten besser zu berücksichtigen [21].

Die Berücksichtigung von Concept Drift ist entscheidend, um die langfristige Leistung und Zuverlässigkeit von maschinellen Lernmodellen sicherzustellen. Durch die Verwendung von Techniken zur Vermeidung von Concept Drift können Modelle an sich ändernde Bedingungen angepasst werden und eine kontinuierlich gute Leistung erzielen. In Kapitel 4 wird bezogen auf die hier in dieser Arbeit verwendeten Daten erneut auf Concept Drift eingegangen und erläutert, in welcher Form Concept Drift in diesen auftritt.

3.3 Baumbasierte Lernalgorithmen

Baumbasierte Lernalgorithmen stellen eine bedeutende Klasse von Algorithmen im Bereich des maschinellen Lernens dar und haben in den letzten Jahren erhebliche Aufmerksamkeit und Anwendung in verschiedenen Domänen erhalten. In diesem Abschnitt wird auf den Aufbau und die Funktionsweise von zwei baumbasierten Lernalgorithmen eingegangen: „Decision Trees“ und „Random Forests“. In den Studien zu Churn Predictions in dem Abschnitt 2 zu Verwandten Arbeiten haben sich baumbasierte Lernalgorithmen als beste Wahl für den Anwendungsfall Churn Predictions bewiesen. Aufgrund dessen wird diese Arbeit keine erneute Evaluierung von Klassifikationsalgorithmen durchführen, sondern sich nur auf baumbasierte Lernalgorithmen und Neuronale Netze fokussieren. Anzumerken ist, dass „Gradient Boosting“ ebenfalls ein oft erwähnter Lernalgorithmus in den verwandten Arbeiten war. Dieser aufgrund von technischen Hindernissen, innerhalb der virtuellen Maschine welche exklusiv den Zugriff auf die hier verwendeten Daten ermöglicht, nicht untersucht wird. Hier kann erneut angemerkt werden, dass die Wahl der korrekten Features einen signifikant größeren Einfluss auf die Ergebnisse hat als der Klassifikator selbst. Die Unterschiede in den Ergebnissen zwischen Gradient Boosting und Random Forest waren ebenfalls maginal [30].

3.3.1 Decision Trees

Decision Trees sind einfache, aber äußerst wirkungsvolle Modelle, die eine hierarchische Struktur in Form eines binären oder mehrwegigen Baums aufweisen. Aufgrund ihrer Fähigkeit, leicht interpretierbar zu sein und klare Entscheidungswege zu bieten, haben sie zahlreiche Anwendungen in Klassifikations- und Regressionsaufgaben gefunden. Ihre Effizienz und Einfachheit machen sie zu einem beliebten Werkzeug in der Datenanalyse und dem maschinellen Lernen.

Entscheidungsregeln und Partitionierung:
Der Aufbau eines Decision Trees erfolgt durch das rekursive Partitionierungsverfahren,

welches den Datensatz in Untergruppen aufteilt, um homogene Subsets zu erzeugen. Der Entscheidungsprozess beginnt an der Wurzel des Baums, wo das Merkmal ausgewählt wird, das den größten Informationsgewinn oder die größte Reduktion des Gini- oder Entropy-Index ermöglicht. Dieser Schritt ist entscheidend für die Auswahl des Merkmals, das den besten Split der Daten ermöglicht, und ermöglicht die bestmögliche Trennung der Zielvariablen in den entstehenden Untergruppen. Die Daten werden dann entsprechend der Bedingungen des ausgewählten Merkmals in zwei oder mehr Untergruppen geteilt. Dieser Vorgang wird für jede Untergruppe wiederholt, indem für jede neu entstandene Gruppe das beste Trennmerkmal identifiziert wird. Dies führt zur schrittweisen Partitionierung der Daten und bildet die Hierarchie des Entscheidungsbaums. Der Prozess wird solange wiederholt, bis eine bestimmte Stoppbedingung erreicht ist, die normalerweise durch eine maximale Baumtiefe oder eine Mindestanzahl von Datenpunkten in den Blattknoten festgelegt wird.

Interpretierbarkeit und Anwendungen:
Die Interpretierbarkeit von Decision Trees ist ein entscheidender Vorteil, da die Entscheidungskriterien und Regeln klar und nachvollziehbar sind. Dies macht sie besonders geeignet für Anwendungen, in denen die Erklärbarkeit der Entscheidungsgrundlagen wichtig ist, wie beispielsweise im medizinischen Bereich oder bei Finanzanalysen. Entscheidungsbäume finden Anwendung in einer Vielzahl von Bereichen, einschließlich der medizinischen Diagnose, dem Customer Relationship Management (CRM), der Betrugserkennung, der Kreditvergabe und vielen anderen Domänen. In Klassifikationsaufgaben ermöglichen sie die Zuordnung von Datenpunkten zu bestimmten Klassen oder Kategorien, während sie in Regressionsaufgaben kontinuierliche Werte vorhersagen können. Angesichts ihrer einfachen Struktur und Interpretierbarkeit sind Decision Trees ein wertvolles Werkzeug in der Datenanalyse, das eine effektive Lösung für verschiedene Probleme bietet und gleichzeitig Einblicke in die Entscheidungsfindung des Modells ermöglicht.

3.3.2 Random Forest

Random Forest ist ein Ensemble-Lernalgorithmus, der auf der Idee der Kombination mehrerer Decision Trees basiert. Durch die Zusammenführung von Entscheidungsbäumen strebt Random Forest danach, die Leistungsfähigkeit und Robustheit des Modells zu verbessern, indem es die Vorhersagegenauigkeit erhöht und gleichzeitig das Risiko von Überanpassung (Overfitting) reduziert.

Aufbau eines Random Forests:
Der Aufbau eines Random Forests besteht aus zwei entscheidenden Hauptkomponenten: der zufälligen Auswahl von Datenpunkten und der zufälligen Auswahl von Merkmalen. Zunächst wird für jeden Entscheidungsbaum im Random Forest ein Unterdatensatz aus dem ursprünglichen Datensatz durch zufälliges Sampling mit Zurücklegen (Bootstrapping) erstellt. Dies bedeutet, dass für jeden Baum eine Teilmenge von Datenpunkten aus dem Gesamtdatensatz gezogen wird. Dieser Prozess ermöglicht es, dass einige Datenpunkte in einem Baum mehrfach und andere gar nicht verwendet werden können, wodurch die Variabilität und Robustheit des Modells erhöht wird.

Darüber hinaus wird für jeden Baum nur eine zufällige Teilmengen von Merkmalen aus dem Gesamtsatz ausgewählt, um den Entscheidungsprozess durchzuführen. Diese Technik wird als "Feature-Bagging"bezeichnet und führt dazu, dass jeder Baum auf einem unterschiedlichen Untermengen der Merkmale trainiert wird. Dies erhöht die Dekorrelation zwischen den Bäumen und ermöglicht eine bessere Berücksichtigung verschiedener relevanter Merkmale.

Vorhersage und Ensemble-Bildung:
Nach dem Training jedes Entscheidungsbaums im Random Forest erfolgt die Vorhersage für neue Datenpunkte durch eine Mehrheitsentscheidung der Einzelvorhersagen der Bäume. Dies bedeutet, dass die finale Vorhersage des Random Forests durch eine Mehrheitsabstimmung oder eine Durchschnittsbildung der Vorhersagen der einzelnen Bäume erfolgt. Diese Aggregation kombiniert die Stärken der einzelnen Bäume und führt zu einer robusten und leistungsstarken Vorhersagefähigkeit des Modells. Durch die Kombination von mehreren Entscheidungsbäumen ermöglicht Random Forest eine verbesserte Modellleistung, die in vielen Anwendungen von großem Nutzen ist. Der Algorithmus reduziert die Gefahr von Overfitting und verbessert die Generalisierungsfähigkeit, wodurch er sich insbesondere für komplexe Probleme und große Datensätze eignet. Random Forest hat sich als eine robuste und weit verbreitete Methode erwiesen, die in einer Vielzahl von Domänen erfolgreich eingesetzt wird.

Vergleich zwischen Decision Trees und Random Forests

Eigenschaften	Decision Trees	Random Forests
Interpretierbarkeit	Einfach interpretierbar, klare Entscheidungsregeln	Weniger leicht interpretierbar, Ensemble-Modell mit aggregierten Entscheidungen
Overfitting	Tendenz zu Overfitting, hohe Varianz	Reduziert Overfitting, verbesserte Generalisierungsfähigkeit durch Ensemble-Techniken
Leistungsfähigkeit	Eingeschränkte Präzision, einfache Probleme	Höhere Präzision, komplexe Probleme, große Datensätze
Rechenzeit	Schneller Aufbau	Höhere Rechenzeit wegen Ensemble-Bildung

Tabelle 3.1: Vergleich zwischen Decision Trees und Random Forests

Insgesamt zeigen Decision Trees eine einfache Struktur und Interpretierbarkeit, sind jedoch anfälliger für Overfitting und liefern in komplexen Problemen möglicherweise nicht die beste Leistung. Random Forests hingegen verbessern die Vorhersagegenauigkeit und Robustheit durch Ensemble-Techniken, aber sie sind weniger interpretierbar und erfordern etwas mehr Rechenzeit. Die Wahl zwischen diesen Algorithmen hängt von den spezifischen Anforderungen und dem gewünschten Kompromiss zwischen Präzision und Interpretierbarkeit ab.

3.4 Enkodierung von kategorialen Variablen

Das One-Hot Encoding ist eine weit verbreitete Technik in der Datenverarbeitung und im maschinellen Lernen, um kategoriale Daten numerisch darzustellen. Oftmals enthalten Datensätze Merkmale, die nicht direkt in numerischer Form vorliegen, sondern in Form von Kategorien oder Labels, z. B. Farben, Ländernamen oder Berufsklassen. Diese kategorialen Daten können von vielen maschinellen Lernalgorithmen nicht direkt verarbeitet werden, da diese in der Regel nur mit numerischen Werten arbeiten. One-Hot Encoding ist eine Methode, um solche kategorialen Merkmale in binäre Vektoren umzuwandeln, die von maschinellen Lernalgorithmen verarbeitet werden können. Da der hier verwendete Datensatz viele kategoriale Merkmale enthält soll diese Technik der Kodierung in diesem Kapitel kurz erläutert werden.

Funktionsweise des One-Hot Encoding

Die grundlegende Idee des One-Hot Encoding besteht darin, jedes kategoriale Merkmal in eine neue Menge von binären Merkmalen umzuwandeln, wobei jedes binäre Merkmal nur einen möglichen Wert des ursprünglichen kategorialen Merkmals darstellt. Angenommen, wir haben ein kategoriales Merkmal Farbe mit den möglichen Werten „Rot“, „Blau“ und „Grün“. Beim One-Hot Encoding wird dieses Merkmal in drei neue binäre Merkmale umgewandelt: „Farbe_Rot“, „Farbe_Blau“ und „Farbe_Grün“. Jedes dieser binären Merkmale kann nur den Wert 0 oder 1 annehmen, je nachdem, ob der ursprüngliche Wert des kategorialen Merkmals mit dem Wert des jeweiligen binären Merkmals übereinstimmt.

Farbe	Farbe_Rot	Farbe_Blau	Farbe_Grün
Rot	1	0	0
Blau	0	1	0
Grün	0	0	1

Tabelle 3.2: Beispiel eines One-Hot Encoding für Farben

Jede Zeile des ursprünglichen Datensatzes wird in eine separate Zeile im neuen One-Hot-Encoding-Datensatz umgewandelt, wobei jedes binäre Merkmal den Wert 0 oder 1 hat, um die jeweilige Farbe darzustellen.

Anwendungsfälle des One-Hot Encoding

One-Hot Encoding wird häufig in maschinellen Lernanwendungen eingesetzt, bei denen kategorialen Merkmale in numerische Form umgewandelt werden müssen, um von den Lernalgorithmen verarbeitet werden zu können. Beispiele für solche Anwendungsfälle sind:

- **Klassifikation:** Bei der Klassifikation von Daten in verschiedene Klassen ist es oft erforderlich, kategoriale Merkmale in numerische Form zu überführen, damit die Klassifikationsalgorithmen sie verarbeiten können.

- **Clustering:** Beim Clustering von Daten werden ähnliche Daten in Gruppen zusammengefasst. Hier kann One-Hot Encoding verwendet werden, um kategoriale Merkmale in einem numerischen Raum darzustellen, um Ähnlichkeiten zwischen den Daten besser zu erfassen.
- **Feature Engineering:** Beim Feature Engineering werden neue Merkmale aus den vorhandenen Daten abgeleitet, um die Leistung der Lernalgorithmen zu verbessern. One-Hot Encoding kann dabei helfen, kategoriale Merkmale in geeignete numerische Formen zu überführen, die für die Modellierung relevant sind.

3.5 Evaluation von maschinellen Lernmodellen

Die Auswahl des richtigen maschinellen Lernmodells ist ein entscheidender Schritt bei der Lösung von Problemen im maschinellen Lernen. Es gibt eine Vielzahl von Modellen mit unterschiedlichen Eigenschaften, Fähigkeiten und Anwendungsgebieten. In diesem Kapitel werden verschiedene Techniken zur Auswahl von maschinellen Lernmodellen erläutert, um eine fundierte Entscheidung zu treffen.

3.5.1 Modellbewertungsmetriken

Die Bewertung der Leistung von maschinellen Lernmodellen erfordert die Verwendung von Metriken, um die Vorhersagequalität und die Genauigkeit der Modelle zu bewerten.

Eine grundlegende Metrik zur Bewertung von Klassifikationsmodellen ist die Genauigkeit (Accuracy). Sie gibt den Anteil der korrekten Vorhersagen im Verhältnis zur Gesamtzahl aller Instanzen im Datensatz an. Die Genauigkeit ist eine einfache und intuitive Metrik, die jedoch in bestimmten Szenarien irreführend sein kann, insbesondere wenn die Klassen im Datensatz unausgewogen sind. Präzision (Precision) und Recall sind Metriken, die häufig in Verbindung mit der Genauigkeit verwendet werden. Präzision gibt an, wie viele der positiven Vorhersagen tatsächlich korrekt sind, während Recall angibt, wie viele der tatsächlich positiven Instanzen richtig erkannt wurden. Präzision und Recall sind insbesondere bei unausgewogenen Klassenverteilungen wichtig, um die Leistung des Modells für bestimmte Klassen genauer zu bewerten. Der F1-Score ist eine Metrik, die das Gleichgewicht zwischen Präzision und Recall herstellt. Er kombiniert diese beiden Metriken zu einer einzigen Kennzahl, die die Mittelwerte von Präzision und Recall berechnet. Der F1-Score ist besonders nützlich, wenn Präzision und Recall gleichermaßen wichtig sind. Eine weitere wichtige Metrik ist die AUC-ROC (Area Under the Receiver Operating Characteristic Curve). Sie wird häufig für binäre Klassifikationsmodelle verwendet und misst die Fähigkeit des Modells, zwischen den Klassen zu unterscheiden. Eine AUC-ROC von 1 zeigt ein perfektes Modell an, während ein Wert von 0,5 auf ein zufälliges Modell hinweist. Neben diesen grundlegenden Metriken gibt es auch spezifischere Metriken, die je nach Anwendungsgebiet und Anforderungen des Problems verwendet werden können. Beispiele dafür sind Sensitivität, Spezifität, logarithmischer Verlust (log loss) und Gini-Koeffizient.

In den folgenden Abschnitten dieser Arbeiten werden diese Begrifflichkeiten regelmäßig verwendet, um verschiedene Aspekte der Analyse und Modellevaluation zu diskutieren und die zugrunde liegenden Konzepte zu vertiefen.

3.5.2 Holdout Verfahren

Das Holdout Verfahren in der Datenanalyse, auch als Train-Test-Split-Verfahren bezeichnet, wird häufig für das Training und die Evaluation von maschinellen Lern-Modellen verwendet. Es ist ein einfaches, aber dennoch effektives Verfahren, um die Leistung eines Modells auf unbekannten Daten abzuschätzen. Das Holdout Verfahren ist eine Form der Kreuzvalidierung, bei der der verfügbare Datensatz in zwei disjunkte Teilmengen aufgeteilt wird: eine Trainingsmenge und eine Testmenge. Die Trainingsmenge wird verwendet, um das Modell zu trainieren, während die Testmenge dazu dient, die Leistung des Modells zu bewerten, indem es auf unbekannten Daten getestet wird.

Es wird in verschiedenen Szenarien angewendet. Erstens dient es dazu, die Leistung eines trainierten Modells objektiv zu bewerten, bevor es auf neue, nicht gesehene Daten angewendet wird. Dadurch wird verhindert, dass das Modell die Testdaten „auswendig lernt“ und die Leistungsfähigkeit auf unbekannten Daten übertrieben einschätzt. Zweitens kann das Holdout Verfahren beim Hyperparameter-Tuning verwendet werden. Hierbei werden verschiedene Kombinationen von Hyperparametern verglichen, um diejenigen auszuwählen, die die beste Leistung auf der Testmenge zeigen. Drittens kann es bei der Feature-Auswahl eingesetzt werden, um die relevantesten Merkmale zu identifizieren. Man trainiert das Modell mit verschiedenen Feature-Sets und vergleicht deren Leistung auf der Testmenge, um die besten Merkmale auszuwählen.

Das Vorgehen beim Holdout Verfahren besteht aus mehreren Schritten. Zunächst wird der verfügbare Datensatz zufällig in zwei Teilmengen aufgeteilt: eine Trainingsmenge und eine Testmenge. Üblicherweise werden etwa 70-80% der Daten für das Training und der Rest für das Testen verwendet. Anschließend wird das Machine Learning-Modell auf der Trainingsmenge trainiert, indem es aus den Daten die Zusammenhänge zwischen den Eingangsvariablen und den Zielvariablen erfasst. Danach erfolgt die Bewertung des Modells auf der Testmenge, wobei die Leistung anhand von geeigneten Metriken wie Genauigkeit, F1-Score oder anderen Fehlermaßen bewertet wird. Die auf der Testmenge erhaltene Leistung dient als Abschätzung der Leistung des Modells auf unbekannten Daten. Dabei sollte die Testmenge nur einmal verwendet werden, um eine unverfälschte Bewertung des Modells zu gewährleisten.

3.5.3 Kreuzvalidierung

Kreuzvalidierung (Cross Validation) ist eine bewährte Methode zur Bewertung der Leistung von maschinellen Lernmodellen. Es ermöglicht die Schätzung der Leistung eines Modells auf der Grundlage eines begrenzten Datensatzes, wodurch mögliche Probleme wie Überanpassung vermieden werden können. In diesem Abschnitt werden die wichtigsten Aspekte der Kreuzvalidierung erläutert.
Ein grundlegendes Konzept der Kreuzvalidierung besteht darin, den verfügbaren

Datensatz in mehrere Teile aufzuteilen, wobei ein Teil für das Training des Modells verwendet wird und die restlichen Teile für die Evaluation der Leistung verwendet werden. Eine häufig verwendete Variante ist die k-fache Kreuzvalidierung, bei der der Datensatz in k gleich große Teile aufgeteilt wird. Das Modell wird dann k Mal trainiert und evaluiert, wobei in jedem Durchlauf ein anderer Teil als Testdatensatz verwendet wird. Ein wichtiger Aspekt der Kreuzvalidierung ist die Berücksichtigung der Datenaufteilung. Um die Ergebnisse robuster zu gestalten, wird oft die stratifizierte Kreuzvalidierung verwendet. Dabei wird sichergestellt, dass die Verteilung der Zielvariablen in den einzelnen Teilen des Datensatzes ähnlich ist. Dies ist besonders wichtig, wenn die Zielvariable eine ungleichmäßige Verteilung aufweist.

Die Leistung des Modells wird während der Kreuzvalidierung anhand verschiedener Metriken bewertet. Häufig verwendete Metriken sind Genauigkeit, Präzision, Recall, F1-Score und AUC-ROC. Diese Metriken ermöglichen eine quantitative Bewertung der Vorhersagequalität und helfen dabei, Modelle zu vergleichen und die besten Ergebnisse zu erzielen. Kreuzvalidierung bietet mehrere Vorteile. Erstens ermöglicht es eine zuverlässige Schätzung der Leistung eines Modells, indem es eine bessere Nutzung der vorhandenen Daten ermöglicht. Zweitens hilft es, Überanpassung zu vermeiden, indem es das Modell auf Daten evaluiert, die es nicht während des Trainings gesehen hat. Drittens ermöglicht es die Identifizierung von Modellen, die besser auf unbekannte Daten generalisieren können [31].

3.5.4 Modellvergleich und -auswahl

Der Modellvergleich und die Modellauswahl spielen eine entscheidende Rolle bei der Auswahl des richtigen maschinellen Lernmodells. Verschiedene Aspekte müssen berücksichtigt werden, um eine fundierte Entscheidung zu treffen. Bei der Bewertung und dem Vergleich verschiedener maschineller Lernmodelle sind Bewertungskriterien von großer Bedeutung. Bewertungsmetriken ermöglichen die quantitative Bewertung der Leistung und Genauigkeit der Modelle. Die hierfür am häufigsten verwendeten Metriken wurden bereits in dem Unterkapitel Modellbewertungsmetriken aufgelistet und erläutert.

Ein wichtiger Aspekt beim Modellvergleich ist die Berücksichtigung verschiedener Modelle mit unterschiedlichen Eigenschaften und Algorithmen. Es gibt eine Vielzahl von maschinellen Lernmodellen wie Entscheidungsbäume, Support Vector Machines (SVM), k-Nearest Neighbors (k-NN), Random Forests und neuronale Netzwerke, um nur einige zu nennen. Jedes Modell hat seine eigenen Vor- und Nachteile in Bezug auf Genauigkeit, Robustheit, Interpretierbarkeit und Rechenkomplexität. Daher ist es wichtig, verschiedene Modelle aus verschiedenen Klassen zu betrachten und ihre Eignung für das spezifische Problem zu bewerten.

Ein weiterer wichtiger Aspekt ist der Vergleich von Modellarchitekturen. In einigen Fällen kann eine bestimmte Modellarchitektur besser für das gegebene Problem geeignet sein als andere. Der Vergleich von Modellarchitekturen kann durch Experimente und die Bewertung der Leistung anhand ausgewählter Metriken erfolgen. Bei der Modellauswahl spielt auch die Optimierung von Hyperparametern eine wesentliche

Rolle. Jedes maschinelle Lernmodell hat Hyperparameter, die die Leistung des Modells beeinflussen. Die Wahl der optimalen Hyperparameterwerte kann durch Techniken wie Grid-Search, Random-Search oder Bayes'sche Optimierung erfolgen [8]. Durch die Optimierung der Hyperparameter kann die Leistung des Modells maximiert und an die spezifischen Anforderungen des Problems angepasst werden. Die Auswahl des besten Modells erfolgt durch einen umfassenden Vergleich der Leistung, des Modellverhaltens und der Anforderungen des Problems. Neben der quantitativen Bewertung der Metriken spielen auch qualitative Aspekte wie die Interpretierbarkeit des Modells, die Robustheit gegenüber Ausreißern und die Skalierbarkeit eine Rolle. Zusammenfassend kann gesagt werden, dass der Modellvergleich und die Modellauswahl eine umfassende Bewertung verschiedener maschineller Lernmodelle erfordern. Dabei sollten Bewertungskriterien, der Vergleich von Modellarchitekturen und die Optimierung von Hyperparametern berücksichtigt werden.

Kapitel 4

Datenbeschaffung und - vorbereitung

Die vorliegende Arbeit nutzt als primäre Datengrundlage einen Datensatz des Partnerunternehmens, welches Daten zu circa 5 Mio. Arbeitnehmern und ihren Gehältern pro Monat enthält. Diese sind eine nicht repräsentative Stichprobe der 14.5 Mio. monatlichen Lohnabrechnungen. In diesem Kapitel soll nun der Entstehungsprozess von Lohnabrechnungsdaten bis zu den finalen Features beschrieben werden.

4.1 Datenvorbereitung

Im Rahmen dieser Arbeit erfolgt u.a. eine Eingrenzung des Datensatzes auf:

- Sozialversicherungspflichtig Beschäftigte
- Arbeitnehmer zwischen 18 und 60 Jahren
- Keine geringfügig Beschäftigten
- Keine Daten aus Branchsektoren T und U (vgl. Tabelle 4.1)
- Keine Daten aus dem öffentlichen Dienst, Baulohn sowie zu Geschäftsführern

Das Hauptziel dieser Arbeit ist die Vorhersage von Kündigungen, wobei keine Unterscheidung zwischen den verschiedenen Arten von Abgängen möglich ist. Dies resultiert daraus, dass eine „Kündigung“ allein durch das Feld „exit“ bestimmt werden kann welches in der Lohnabrechnungssoftware gesetzt wird. Bei diesem Feld handelt es sich um ein Austrittsdatum. Um den Faktor „Rentenaustritte“ auszuschließen, wurde eine Obergrenze von 60 Jahren gewählt. Allerdings stellt sich die Herausforderung, dass keine Unterscheidung zwischen den verschiedenen Arten von Kündigungen vorgenommen werden kann. Die Arten unterscheiden sich zwischen objektiven Kündigungen, subjektiven Kündigungen sowie Kündigungen seitens des Arbeitgebers. Zusätzlich wurden geringfügig Beschäftigte und nicht festangestellte Arbeitnehmer, wie beispielsweise Auszubildende und Werkstudenten, von der Datenauswahl ausgeschlossen. Dies geschah, da diese Gruppen nicht zu unserer definierten Zielgruppe gehören. Die verfügbaren Daten enthalten Informationen zu verschiedenen Branchen-

sektoren, die wiederum die verschiedenen Wirtschaftszweige in einzelne Kategorien unterteilen. Diese Unterteilung basiert auf dem Klassifikationsbericht des Statistischen Bundesamts [14]. Die beiden Sektoren „T“ und „U“ werden dabei herausgefiltert, da sie zu einer bestimmten Nische gehören und nicht Zielgruppe der Kündigungsvorhersage sein sollen. Eine detaillierte Beschreibung der einzelnen Sektoren findet sich in Tabelle 4.1. Zuletzt werden Personen aus dem öffentlichen Dienst, dem Baulohn, sowie Geschäftsführer herausgefiltert.

Sektoren	**Beschreibung**
A	Land- und Forstwirtschaft, Fischerei
B	Bergbau und Gewinnung von Steinen und Erden
C	Verarbeitendes Gewerbe
D	Energieversorgung
E	Wasserversorgung; Abwasser- und Abfallentsorgung und Beseitigung von Umweltverschmutzungen
F	Baugewerbe
G	Handel; Instandhaltung und Reparatur von Kraftfahrzeugen
H	Verkehr und Lagerei
I	Gastgewerbe
J	Information und Kommunikation
K	Erbringung von Finanz- und Versicherungsdienstleistungen
L	Grundstücks- und Wohnungswesen
M	Erbringung von freiberuflichen, wissenschaftlichen und technischen Dienstleistungen
N	Erbringung von sonstigen wirtschaftlichen Dienstleistungen
O	Öffentliche Verwaltung, Verteidigung; Sozialversicherung
P	Erziehung und Unterricht
Q	Gesundheits- und Sozialwesen
R	Kunst, Unterhaltung und Erholung
S	Erbringung von sonstigen Dienstleistungen
T	Private Haushalte mit Hauspersonal; Herstellung von Waren und Erbringung von Dienstleistungen durch Private Haushalte für den Eigenbedarf ohne ausgeprägten Schwerpunkt
U	Exterritoriale Organisationen und Körperschaften

Tabelle 4.1: Beschreibung der verschiedenen Branchsektoren. [14]

Die verbleibenden Einträge der Arbeitnehmer in den vorliegenden Daten enthalten eine Vielzahl von Informationen, die eine breite Palette abdecken. Diese Informationen umfassen beispielsweise das Alter, das Gehalt, die wöchentliche Arbeitszeit, den Beruf

und die steuerlichen Abgaben, um nur einige zu nennen. Diese Daten sind bereits teilweise vorverarbeitet worden, indem sie aus den Lohnabrechnungen extrahiert und als separate Merkmale in einem Datensatz gespeichert wurden. Im folgenden Abschnitt „Feature-Engineering“ wird genauer auf die erforderlichen Verarbeitungsschritte eingegangen und erläutert, welche Merkmale tatsächlich in das Training eines Lernmodells einfließen. Dabei wird herausgearbeitet, wie diese Merkmale sinnvoll für das Modell genutzt werden können.

4.2 Feature-Engineering

Feature-Engineering ist ein wesentlicher Schritt bei der Vorbereitung von Daten für maschinelles Lernen. Es umfasst die Schaffung und Auswahl relevanter Merkmale (auch als Eingangsvariablen oder Prädiktoren bezeichnet) aus den Rohdaten, um die Leistung und Vorhersagekraft des Modells zu verbessern. Neben den bereits in Abschnitt 4.1 erwähnten Filterungen sind noch weitere Feature-Engineering Schritte durchgeführt wurden. Technisch wurde hier Python in Kombination mit PySpark SQL und Pandas verwendet.

4.2.1 Eingabevariablen

Das erste zusätzlich konstruierte Feature ist die Distanz zwischen Arbeitnehmer und Arbeitgeber (employee to employer distance). Diese Distanz zwischen den Koordinaten (Breitengrad lat_1, Längengrad lon_1) und (Breitengrad lat_2, Längengrad lon_2) kann durch die Funktion distance$(lat_1, lon_1, lat_2, lon_2)$ (Haversine-Formel) berechnet werden:

$$\begin{aligned}\text{distance}(lat_1, lon_1, lat_2, lon_2) = 6371.0 \times \arccos(\\ \sin(\frac{\pi \cdot lat_1}{180}) \cdot \sin(\frac{\pi \cdot lat_2}{180}) + \cos(\frac{\pi \cdot lat_1}{180}) \cdot \cos(\frac{\pi \cdot lat_2}{180}) \cdot \cos(\frac{\pi \cdot (lon_1 - lon_2)}{180}))\end{aligned} \tag{4.1}$$

Das nächste Merkmal ist die Gehaltsmarktwertprognose (yearly predicted gross payment), die mithilfe eines Regressionsmodells generiert werden kann, welches bereits bei PBo verwendet wird. Über diesen Wert kann nun der Gehaltsunterschied (salary difference) zwischen Markwertprognose und Jahresgehalt berechnet werden.

Ein weiteres Merkmal ist die Fluktuationsrate je Unternehmen. Die Fluktuationsrate wird als ein Feature berücksichtigt, das die Mitarbeiterfluktuation innerhalb jedes Unternehmens widerspiegelt. Die Periode kann abhängig des Anwendungsfalls gewählt werden, hier ist diese auf ein Jahr festgelegt. Diese lässt sich wie folgt über die sogenannte „Schlüter-Formel“ berechnen:

$$Fluktuationsrate = \frac{\text{AbgaengeWaehrendPeriode}}{\text{PersonalbestandBeginnPeriode} + \text{ZugaengeWaehrendPeriode}} \times 100\% \quad (4.2)$$

Zwei weitere Merkmale umfassen die gesamte Anzahl unbezahlter Abwesenheitstage in den letzten sechs Monaten (total absence days), die für jeden Abrechnungsmonat einzeln aufsummiert werden, sowie die gesamte Gehaltserhöhung in den letzten zwölf Monaten (salary raise past 12 extrapolated), die ebenfalls pro Abrechnungsmonat aufsummiert wird. Diese Zeiträume wurden gewählt, da die historischen Daten für einige Arbeitnehmer nicht weiter in die Vergangenheit reichen. Die fehlenden Werte wurde durch den Median des jeweiligen Arbeitnehmers ersetzt. Dieser Ansatz wurde gewählt, um eine übermäßige Verfälschung der Daten durch zu viele Füllwerte zu verhindern.

Eine vollständige Auflistung aller Features ist in Tabelle 4.2 zu finden.

4.2.2 Zielvariable

Das wichtigste Element innerhalb eines binären Supervised-Learning-Problems, ist die Zielvariable (Target Variable). Es stellt sich die grundlegende Frage, wie Vorhersagen für die Zukunft getroffen werden können, wenn die dafür benötigten Informationen noch nicht bekannt sein können. Hier kommt das Prinzip des „Learning From The Past“ [40] zum Einsatz. Der in dieser Arbeit verwendete Ansatz ist in Abbildung 4.1 dargestellt.

Abbildung 4.1: Ein Beispiel zur bestimmung eines Targets für Zukunfsvorhersagen einer Zeitspanne von 6 Monaten.

In diesem Beispiel ist der aktuellste Monat der April 2023. Somit ist es nicht möglich die aktuellen Daten aus diesen Monat herauszuziehen und die Zielvariable sechs Monate in die Zukunft zu definieren. Aufgrund dessen werden die Daten der grundlegenden Features um sechs Monate in die Vergangenheit geschoben. Nun kann die Zielvariable in diesem Fall „exit “, sechs Monate in der Zukunft ermittelt werden. Beachtet man nur die Zielvariable aus dem April 2023, wäre dies eine Vorhersage von Kündigungen genau sechs Monate in der Zukunft. Da unser Ziel jedoch nicht so spezifisch ist und

eine hohe Unausgeglichenheit zwischen Kündigungen und Nicht-Kündigungen vorliegen würde, aggregieren wir nun die sechs Folgemonate des Oktobers 2022, also des Monats, für den die Feature-Daten festgelegt wurden, um eine Zielvariable zu kontruieren. Dementsprechend findet eine Vorhersage von Kündigungen in den folgenden sechs Monaten statt. Dieser Ansatz kann auf verschiedene Zeitspannen angewendet werden. In dieser Arbeit wurden Zeitspannen von drei, sechs und zwölf Monaten evaluiert, wobei sich die prozentualen Anteile von Kündigungen und Nicht-Kündigungen entsprechend unterscheiden. Weitere Details dazu finden sich in Kapitel 6.

Eine vollständige Auflistung aller Features in Tabelle 4.2:

Feature	Beschreibung
age	Alter des Arbeitnehmers
cluster id	Tätigkeitsschlüssel
employee municipality inhabitants	Gemeindeeinwohner – Arbeitnehmer
employee to employer distance	Distanz in Kilometern
employment period	Zeitraum seit initialem Beitritt
gender	Geschlecht
has short time allowance	Kurzarbeitergeld
has short time allowance last year	Kurzarbeitergeld letzes Jahr
level of prof training	Höchste Berufsausbildung
level of education	Höchster Schulabschluss
salary zipcode index	Lohnniveau der Postleizahl
supervisor	Führungsposition oder nicht
total absence days	Abwesenheitstage letzten 6 Mon.
wwh	Vertragliche Wochenstunden
company municipality inhabitants	Gemeindeeinwohner - Arbeitgeber
company size	Mitarbeiteranzahl
fluct rate	Fluktuationsrate des Unternehmens in den letzten 12 Mon.
region	Region - Bundesländer in 4 Gruppen
branch sector	Branchensektor
yearly gross payment extrapolated	Monatsgehalt – 12 Mon.
yearly gross payment normalized extrapolated	Monatsgehalt – 40h – 12 Mon.
salary raise past 12 extrapolated	Gehaltserhöhung – 12 Mon.
salary raise past 12 norm. extrapolated	Gehaltserhöhung – 12 Mon. – 40h
yearly predicted gross payment	Marktprognose jährliches Gehaltes
salary difference	Differenz des jährliches Gehaltes und der Marktprognose
current gross tax normalized extrapolated	Steuerbrutto auf 40h und 12 Mon.
yearly christmas bonus	Jährliches Weihnachtsgeld
yearly employer company pension scheme	Betriebliche Altersvorsorge
yearly holiday bonus	Jährliches Urlaubsgeld
yearly capital formation savings payment	Vermögenswirksame Leistungen
target (exit)	1 = Kündigung, 0 = keine Kündigung

Tabelle 4.2: Beschreibung der Features gruppiert nach Arbeitnehmer, Arbeitgeber, finanziellen Merkmalen und Target

4.2.3 Feature-Bereinigung

Die verschiedenen Features sind nach Konstruieren in unterschiedlichen Pandas bzw. PySpark Dataframes verteilt. Um nun einen einzigen Dataframe zu erstellen müssen diese nun verbunden werden. Hierfür wird die „employee ID“ und „employer ID“ mit entsprechenden SQL JOIN Befehlen verwendet. Es wird darauf geachtet ein LEFT OUTER JOIN zu verwenden, sofern es die gegebene Situation erlaubt, da keine Dateneinträge verloren gehen sollen. Werden beispielsweise die Abwesenheitstage in den letzten 6 Monaten eines Arbeitgebers summiert und dieser hat erst vor 3 Monaten bei dieser entsprechenden Firma angefangen, so würde ein INNER JOIN dafür sorgen, dass der Eintrag dieses Arbeitnehmers herausfällt. Darüber hinaus weisen die Features „level of education“ , „level of prof training “ und „employee to employer distance“ fehlende (NULL-)Werte auf. In der Theorie gibt es verschiedene Strategien, um mit fehlenden Werten in Daten umzugehen. Diese Strategien können je nach Art der Daten und der spezifischen Analyse unterschiedlich effektiv sein. Hier sind einige gängige Ansätze die in diesem Fall verwendet werden könnten:

- **Entfernen von NULL-Werten:** In einigen Fällen können NULL-Werte aus dem Datensatz entfernt werden. Dabei wird jedoch der gesamte Eintrag dieses Arbeitnehmers entfernt. Dies ist jedoch oft nur dann eine geeignete Strategie, wenn der Anteil der NULL-Werte im Vergleich zur Gesamtgröße des Datensatzes vernachlässigbar klein ist.
- **Imputation (Auffüllung):** Imputationsmethoden werden verwendet, um fehlende Werte durch geschätzte Werte zu ersetzen. Es gibt verschiedene Techniken, wie z. B. die Verwendung des Durchschnitts, des Medians oder einer Regression, um fehlende Werte zu ersetzen.
- **Ersatzwert (Sentinel-Wert):** Ersetzung durch einen speziellen Wert wie -1 oder NaN als Ersatzwert, um NULL-Werte zu kennzeichnen.
- **Modellbasierte Imputation:** Hierbei wird ein Modell erstellt, um die fehlenden Werte zu schätzen. Dies kann zum Beispiel mit Hilfe von linearen Regressionen oder anderen ML-Modellen geschehen.
- **Kategorische Variable:** Wenn fehlende Werte in kategorischen Variablen auftreten, kann eine neue Kategorie „unbekannt“ eingeführt werden, um die fehlenden Werte zu kennzeichnen.
- **Hot-Deck-Imputation:** Hierbei werden fehlende Werte durch Werte aus ähnlichen Fällen im Datensatz ersetzt.

In den Merkmalen „level of education“ und „level of prof training“ fehlen etwa 600.000 Werte, wodurch die schlichte Entfernung der NULL-Werte nicht in Frage kommt. Eine Imputation ist hier ebenfalls weniger sinnvoll, da keine klaren Zusammenhänge vorhanden sind, um eine verlässliche Schätzung vorzunehmen. Das Nutzen des Medians als Ersatz könnte eine Option sein, jedoch würde dies die Informationsqualität dieses Merkmals erheblich mindern, da der geschätzte Wert stark von den tatsächlichen Werten abweichen kann. Eine ähnliche Problematik könnte bei der Wahl des häufigsten

Berufs auftreten, was viele falsche Einschätzungen zur Folge hätte.

Zudem gibt es einen starken Zusammenhang zwischen den Merkmalen allerdings fehlen diese in der Regel gemeinsam, wodurch es nicht Möglich ist, durch beispielsweise eine höchste Berufsausbildung „Master“ auf einen höchsten Schulabschluss „Abitur“ zurückzuführen. Aufgrund dieser Bewertungen wurde sich auf die Methode des Sentinel-Wertes festgelegt. Hier wird nun der Wert „-1“ für jeden NULL-Wert eingesetzt. Dies ist möglich, weil einerseits Klassifikatoren wie der Random Forest über die Bibliothek „scikit-learn“ [1], sowie auch Neuronale Netze mit Werten wie diesen umgehen können und diese als „Unbekannte“ klassifizieren. Abschließend werden alle kategorialen Features One-Hot enkodiert nach der Methode aus Abschnitt 3.4. Ebenfalls wird diese Arbeit ein Neuronales Netz verwenden, welches nochmals anders mit den kategorialen Features umgeht. Für dieses werden die Werte nicht One-Hot enkodiert. Aufbau und Erläuterung dieses Netzes folgt in Kapitel 5.

Vorher noch ein detaillierter Einblick in die Daten und deren Zusammenhänge im Abschnitt 4.3.

4.3 Daten-Exploration

Nun soll durch eine Daten-Exploration ein Überblick über den Datensatz und dessen Variablen gegeben werden. Die Variablen werden dabei unter anderem univariat, als auch bivariat betrachtet.

Es werden, wie in Abschnitt 4.1 „Datenvorbereitung“ beschrieben, folgende Daten verwendet:

- Sozialversicherungspflichtig Beschäftigte
- Arbeitnehmer zwischen 18 und 60 Jahren
- Keine geringfügig Beschäftigten
- Keine Arbeitnehmer aus Branchsektoren T und U (vgl. Tabelle 4.1)
- Keine Arbeitnehmer aus öffentlichen Dienst, Baulohn, sowie Geschäftsführer

In Tabelle 4.3 ist eine Übersicht der verschiedenen Variablentypen zusehen:

[1] Scikit-learn.org. (2023): https://scikit-learn.org/stable/

Feature	**Variablentyp**
age	numerisch
current gross tax normalized extrapolated	numerisch
employee municipality inhabitants	numerisch
employee to employer distance	numerisch
employment period	numerisch
salary zipcode index	numerisch
total absence days	numerisch
wwh	numerisch
company municipality inhabitants	numerisch
company size	numerisch
fluct rate	numerisch
salary difference	numerisch
salary raise past 12 extrapolated	numerisch
salary raise past 12 norm. extrapolated	numerisch
yearly capital formation savings payment	numerisch
yearly christmas bonus	numerisch
yearly employer company pension scheme	numerisch
yearly gross payment extrapolated	numerisch
yearly gross payment normalized extrapolated	numerisch
yearly holiday bonus	numerisch
yearly predicted gross payment	numerisch
branch sector	kategorial - nominal
cluster id	kategorial - nominal
region	kategorial - nominal
level of prof training	kategorial - ordinal
level_of_education	kategorial - ordinal
supervisor	boolean
gender	boolean
has short time allowance	boolean
has short time allowance last year	boolean
target	boolean

Tabelle 4.3: Übersicht der Variablentypen

Bruttojahresgehalt

	count	mean	std	25%	50%	75%
Gehalt	1.989.823	35.967€	28.254€	23.447€	31.822€	42.400€

Tabelle 4.4: Zusammenfassung statistischer Werte des Bruttojahresgehaltes

In Tabelle 4.4 ist zu sehen, dass der Durchschnitt des Bruttojahresgehaltes mit 35.967€ über dem 50% Perzentiel (Median) von 31.822€ liegt. Der Countplot in Abbildung 4.2 zeigt eine rechtsschiefe Verteilung des Bruttojahresgehaltes. Es ist wichtig zu beachten, dass es sich hierbei um das nicht normalisierte Bruttojahresgehalt handelt. Dieses schließt alle Arbeitnehmer ein, unabhängig von ihrer Arbeitszeit, sei es Vollzeit oder Teilzeit.

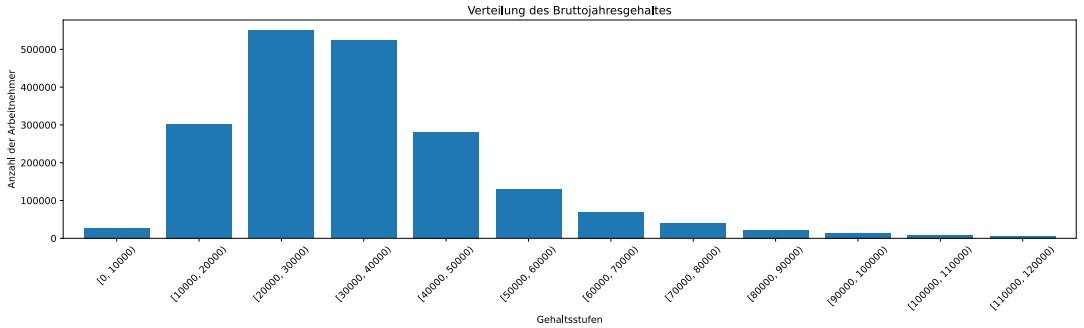


Abbildung 4.2: Verteilung der Jahresgehälter

Das maximale Bruttojahresgehalt beträgt mehrere Millionen Euro. Diese Werte deuten auf das Vorhandensein von Ausreißern in dieser Variablen hin. Da das Ziel dieser Untersuchung jedoch in der Vorhersage von Kündigungen liegt, sind diese Ausreißer relevant und werden beibehalten.

Alter

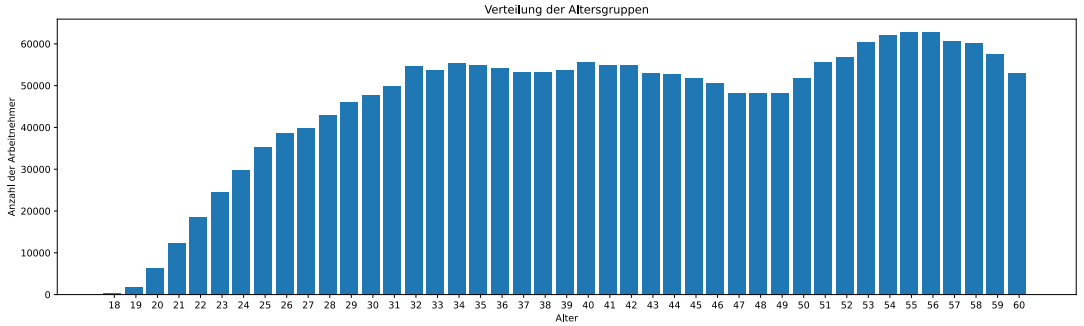


Abbildung 4.3: Verteilung des Alters – univariat

Der Countplot in Abbildung 4.3 veranschaulicht die Verteilung der Arbeitnehmer nach Altersgruppen. Deutlich wird, dass die Anzahl der Arbeitnehmer ab dem 18.

Lebensjahr zunimmt und zwischen dem 30. und 40. Lebensjahr relativ stabil bleibt. Zwischen dem 40. und 49. Lebensjahr zeigt sich ein leichter Abwärtstrend, gefolgt von einem Anstieg bis zum Höhepunkt bei 54 bzw. 55 Jahren. Anschließend nimmt die Anzahl der Arbeitnehmer stetig ab und erreicht bei 60 Jahren, aufgrund der manuellen Eingrenzung, abrupt ein Ende.

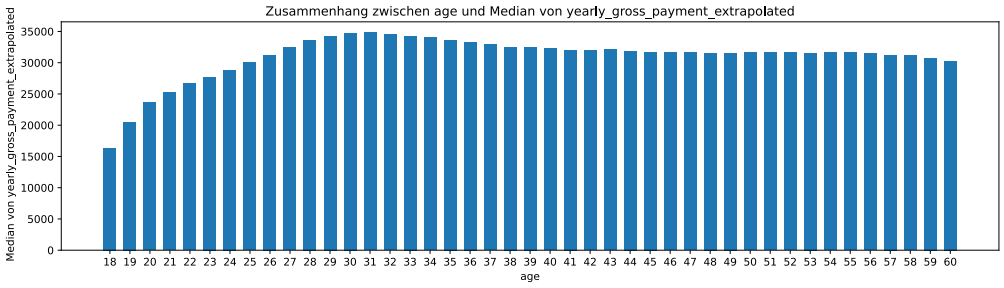


Abbildung 4.4: Bruttojahresgehalt im Bezug zum Alter – bivariat

In Abbildung 4.4 wird die Beziehung zwischen dem Alter und dem Median des Bruttojahresgehalts in einer bivariaten Darstellung veranschaulicht. Hierbei fällt auf, dass das Gehalt von einem Alter von 18 Jahren (ca. 16.000 Euro) bis zum 31. Lebensjahr ansteigt und dabei einen Höhepunkt von etwa 34.000 Euro erreicht. In den darauf folgenden Jahren ist eine leichte Abnahme bis zum 38. Lebensjahr auf etwa 32.000 Euro zu verzeichnen, was auf eine tendenziell höhere Anzahl an Teilzeitbeschäftigungen in dieser Phase zurückzuführen ist. Ab diesem Punkt bleibt der Median bis zum 60. Lebensjahr relativ stabil. Dieses Muster in der Alters-Gehalts-Beziehung kann auf verschiedene Faktoren wie Berufserfahrung, Karriereentwicklung und branchenspezifische Dynamiken zurückzuführen sein.

Geschlecht

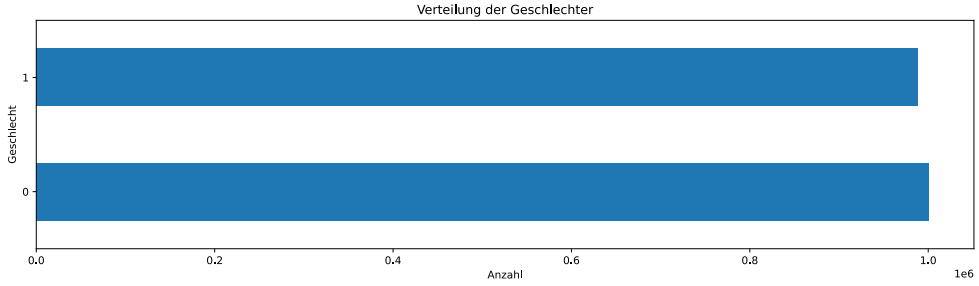


Abbildung 4.5: Verteilung der Geschlechter – univariat

In Abbildung 4.5 wird das univariate Verhältnis zwischen Geschlechtern veranschaulicht. Im vorliegenden Datensatz machen Männer (0) mit einem Anteil von 50,32% und Frauen (1) mit einem Anteil von 49,68% die Geschlechterverteilung aus.

Die bivariate Beziehung zwischen dem Geschlecht und dem Median des Bruttojahresgehalts wird in Abbildung 4.6 anschaulich dargestellt. Dabei fällt auf, dass das Median Bruttojahresgehalt bei Männern (0) mit etwa 37.000 Euro über dem Median

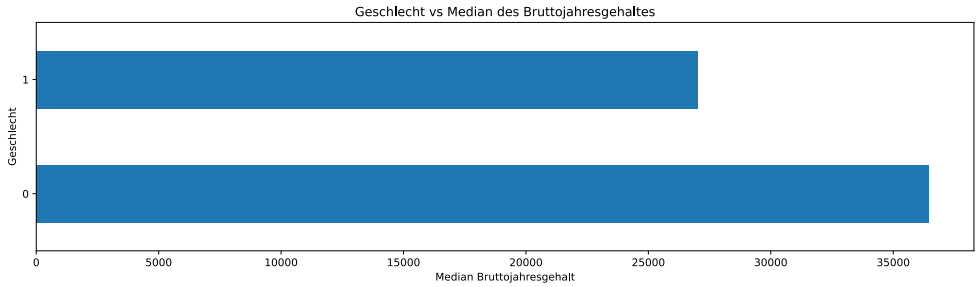


Abbildung 4.6: Bruttojahresgehalt im Bezug zum Geschlecht – bivariat

der Frauen (1) mit ungefähr 27.000 Euro liegt. Wie bereits erwähnt handelt es sich hier um das nicht normalisierte Bruttojahresgehalt. Daher lässt sich diese deutliche Gender-Pay-Gap durch den überproportionalen Anteil von Frauen in Teilzeitarbeit erklären.

Höchster schulischer Bildungsabschluss

Das Merkmal „level of education“ (Höchster Schulabschluss) lässt sich in fünf verschiedene Typen unterteilen:

- Unbekannt (-1)
- ohne Abschluss (1)
- Hauptschulabschluss (2)
- Mittlere Reife (3)
- Abitur (4)

Die Analyse der Schulabschlüsse verdeutlicht, dass der „Mittlere Reife“ Abschluss mit 30,48% am häufigsten vertreten ist, gefolgt von „Unbekannt“ mit 29,49%. Das Abitur weist 23,38% aller Fälle auf, während der Hauptschulabschluss 15,53% ausmacht. Lediglich 1,12% der Arbeitnehmer gehören zur Kategorie „ohne Abschluss“. Diese Verteilung wird auch in Abbildung 4.7 dargestellt.

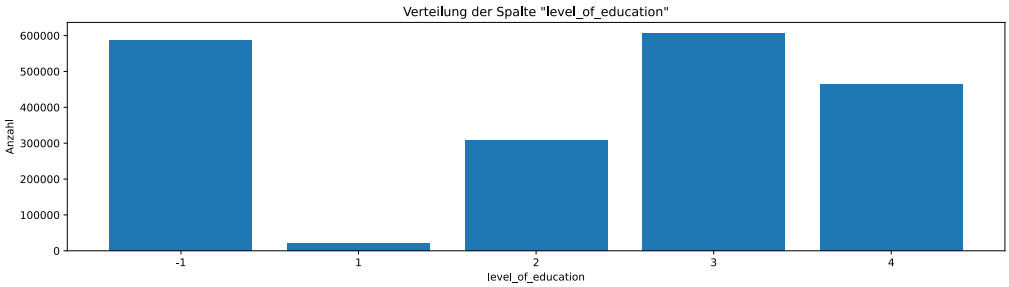


Abbildung 4.7: Verteilung von höchsten Schulabschlüssen – univariat

Bei genauerer Analyse der Gehaltsverteilung nach Schulabschlüssen zeigt sich ein markanter Unterschied im Median-Gehalt. Arbeitnehmer mit Abitur verzeichnen ein um 60% höheres Median-Gehalt (ca. 40.000 Euro) im Vergleich zu Arbeitnehmern ohne Schulabschluss (ca. 25.000 Euro). Die restlichen Kategorien bewegen sich mit etwa 30.000 Euro Median-Gehalt auf ähnlichem Niveau. Dies verdeutlicht auch Abbildung 4.8.

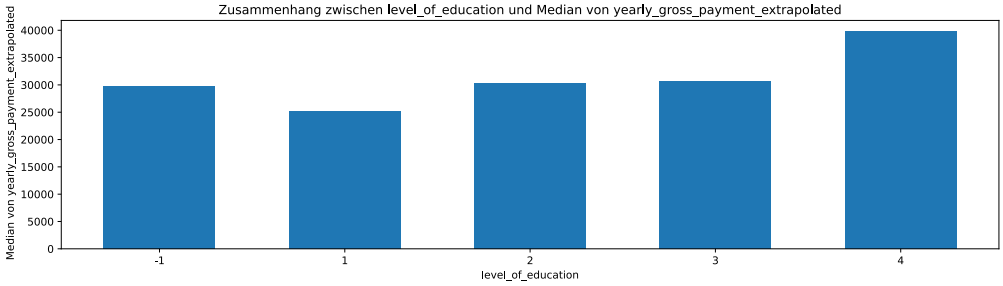


Abbildung 4.8: Bruttojahresgehalt im Bezug zum höchsten Schulabschluss – bivariat

Höchster beruflicher Bildungsabschluss

Das Merkmal „level of prof training“ (Höchster Berufsabschluss) lässt sich in sieben verschiedene Typen unterteilen:

- Unbekannt (-1)
- Ohne Abschluss (1)
- Berufsausbildung (2)
- Meister/Techniker (3)
- Bachelor (4)
- Master/Diplon/Magister (5)
- Promotion (6)

In der Untersuchung der Schulabschlüsse zeigt sich, dass 53,15% der Arbeitnehmer zur Gruppe der „Berufsausbildung“ gehören. Den Status „Unbekannt“ weisen 25,64% der Arbeitnehmer auf, während 7,78% einen „Master/Diplom/Magister“ vorweisen

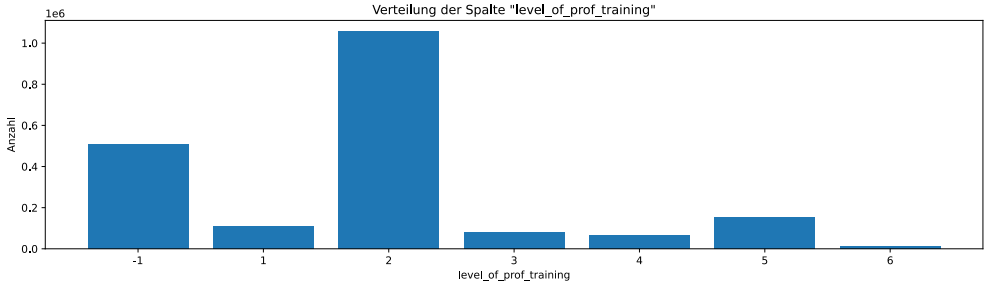


Abbildung 4.9: Verteilung von höchsten Berufsausbildungen – univariat

können. Die Kategorie „Ohne Abschluss“ repräsentiert 5,43% der Daten, gefolgt von 4,14% der Arbeitnehmer mit „Meister/Techniker“ Abschlüssen. 3,24% haben einen Bachelorabschluss erlangt, während die geringste Beteiligung mit 0,60% bei Arbeitnehmern mit Promotion zu verzeichnen ist. (Vgl. Abbildung 4.9)

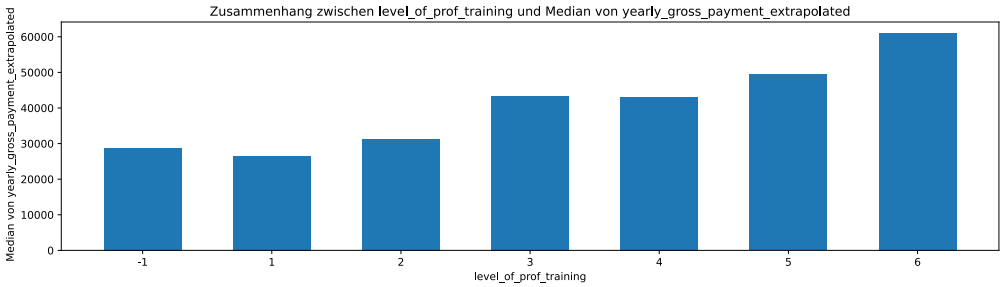


Abbildung 4.10: Bruttojahresgehalt im Bezug zur höchsten Berufsausbildung – bivariat

In Abbildung 4.10 wird ein deutlicher Unterschied im Median-Gehalt dargestellt. Promovierte Arbeitnehmer verzeichnen mit einem Median-Gehalt von etwa 61.000 Euro die höchste Vergütung. Darauf folgen Arbeitnehmer mit einem Masterabschluss mit etwa 49.500 Euro. Die Kategorien Meister/Techniker und Bachelor weisen mit etwa 43.000 Euro eine vergleichbare Vergütung auf. Anschließend liegen Arbeitnehmer mit Berufsausbildung bei etwa 31.000 Euro, gefolgt von der Kategorie „Unbekannt“ mit etwa 29.000 Euro. Arbeitnehmer ohne Abschluss erzielen ein Median-Gehalt von 26.500 Euro.

Beschäftigungsjahre

In der anschließenden univariaten Analyse der Beschäftigungsjahre wird deutlich, dass etwa 40% der Arbeitnehmer eine Beschäftigungsdauer von einem Jahr aufweisen. Ab dem zweiten Beschäftigungsjahr zeigt sich ein konstanter Rückgang der Mitarbeiterzahl.

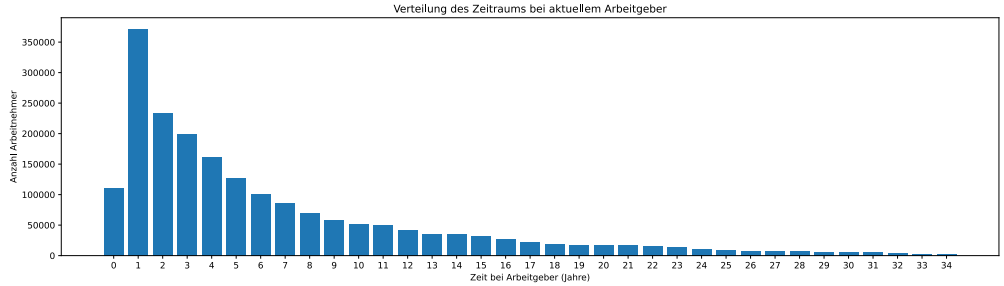


Abbildung 4.11: Verteilung der Beschäftigungsjahre bei aktueller Firma – univariat

Die Beziehung zwischen dem Bruttojahresgehalt und der Beschäftigungsdauer zeigt einen grundsätzlichen, kontinuierlichen Anstieg der Gehälter. Zwischen den Jahren 29 bis 31 ist eine leichte Abnahme erkennbar, die wahrscheinlich auf besondere Gegebenheiten in den Daten zurückzuführen ist und nicht verallgemeinert werden kann. Ebenso zeigt sich eine Abnahme im 46. Beschäftigungsjahr, welche höchstwahrscheinlich auf die geringe Stichprobengröße dieses Jahres zurückzuführen ist.

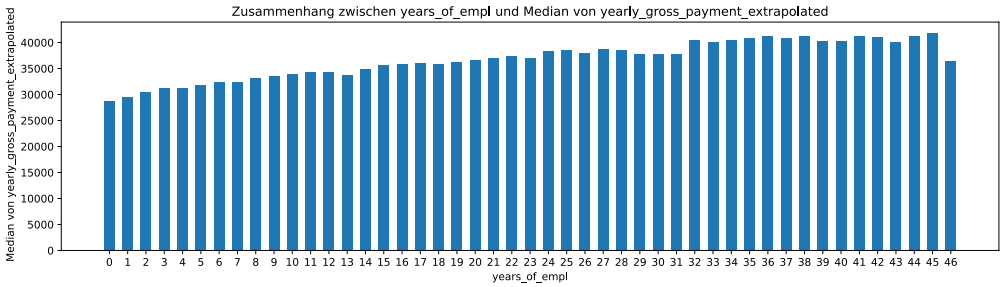


Abbildung 4.12: Bruttojahresgehalt im Bezug zu den Beschäftigungsjahren – bivariat

Unternehmensgröße

In der folgenden Analyse zur Unternehmensgröße wurden die Unternehmen in sieben verschiedene Gruppen unterteilt. Diese reichen von Unternehmen mit bis zu 10 Arbeitnehmern bis hin zu Unternehmen mit mehr als 1001 Arbeitnehmern im Datensatz. Es zeigt sich, dass die Daten eine hohe Anzahl an kleinen und mittleren Unternehmen enthalten. Mit zunehmender Unternehmensgröße nimmt die Anzahl der Arbeitnehmer pro Gruppe ab. Für diese spezifische Datenstichprobe beträgt die durchschnittliche Unternehmensgröße 66, während der Median bei 22 liegt. Um Betriebsgeheimnisse zu schützen, wird auf die Darstellung der univariaten Verteilung der Unternehmensgrößen verzichtet.

Bei dem Verhältnis zwischen dem Bruttojahresgehalt und der Unternehmensgröße zeigt sich ein deutlicher Trend: Größere Unternehmen zahlen im Allgemeinen höhere Gehälter. Dies wird in Abbildung 4.13 veranschaulicht, wo ein kontinuierlicher Anstieg des Median-Gehalts pro Unternehmensgrößen-Kategorie zu beobachten ist.

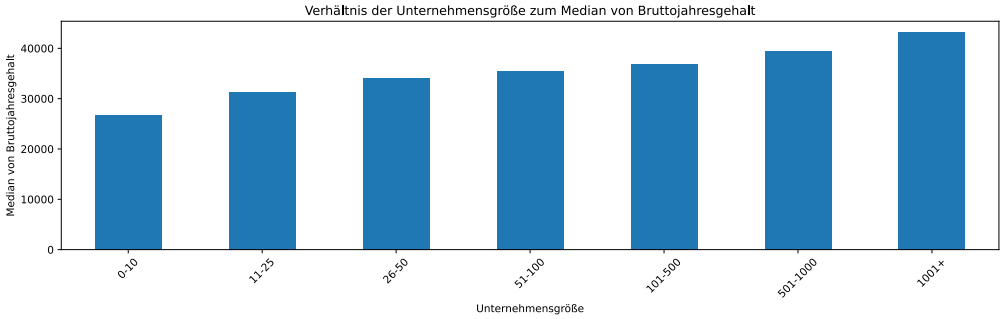


Abbildung 4.13: Bruttojahresgehalt im Bezug zur Unternehmensgröße – bivariat

Branchensektoren

Die univariate Analyse der verschiedenen Branchensektoren, verdeutlicht eine klare Dominanz in einigen Bereichen. Allen voran steht der Sektor „Handel; Instandhaltung und Reparatur von Kraftfahrzeugen“ (G), gefolgt von „Gesundheits- und Sozialwesen“ (Q), „Verarbeitendes Gewerbe“ (C) und „Erbringung von sonstigen wirtschaftlichen Dienstleistungen“ (M). Zur Wahrung von Betriebsgeheimnissen wurde erneut auf die genaue prozentuale Darstellung der univariaten Verteilung der Branchensektoren verzichtet.

Die Analyse der Wechselwirkung zwischen den Branchensektoren und dem Bruttojahresgehalt zeigt einen unterschiedlichen Trend: Hier wird deutlich, dass der Sektor „Information und Kommunikation“ (J) mit einem Median-Gehalt von etwa 50.000 Euro an erster Stelle steht. Auffälliger ist jedoch das untere Ende des Gehaltsspektrums, in dem die Sektoren „Gastgewerbe“ (I) mit ca. 21.000 Euro, „Kunst, Unterhaltung und Erholung“ (R) mit ca. 22.000 Euro und „Erbringung von sonstigen Dienstleistungen“ (S) mit ca. 21.500 Euro liegen. Die übrigen Sektoren zeigen Mediangehälter zwischen ca. 28.000 Euro und 40.000 Euro (siehe Abbildung 4.14).

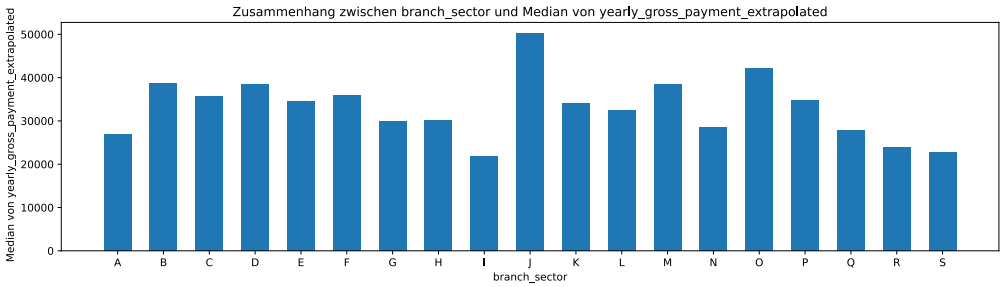


Abbildung 4.14: Bruttojahresgehalt im Bezug zur Branche – bivariat

Region

Das Feature „Region“ basiert auf dem Gemeindeschlüssel des Wohnorts der Arbeitnehmer und klassifiziert diesen in vier geografische Himmelsrichtungen. Die nördliche Region umfasst Bremen, Niedersachsen, Schleswig-Holstein und Hamburg. Die östliche Region umfasst Berlin, Brandenburg, Mecklenburg-Vorpommern, Sachsen-Anhalt, Thüringen und Sachsen. Bayern und Baden-Württemberg gehören zur südlichen Region, während Nordrhein-Westfalen, Hessen, Rheinland-Pfalz und das Saarland zur westlichen Region zählen. Um Betriebsgeheimnisse zu schützen, wird auf die Darstellung der univariaten Verteilung der Unternehmensgrößen verzichtet.

Die Vergütung variiert leicht zwischen den verschiedenen Regionen. In Abbildung 4.15 wird deutlich, dass der Süden ein etwas höheres Median-Gehalt aufweist im Vergleich zu den anderen Regionen. Im Gegensatz dazu zeigt die östliche Region die niedrigste Vergütung innerhalb dieses Datensatzes.

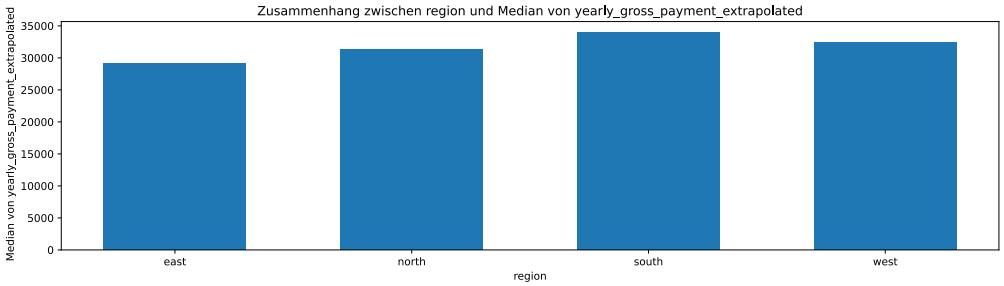


Abbildung 4.15: Bruttojahresgehalt im Bezug zur Region – bivariat

Zusammenhänge zu der Zielvariable (Kündigungen)

Im folgenden Abschnitt werden verschiedene Zusammenhänge zwischen einigen Merkmalen und der Zielvariable analysiert. Dabei liegt der Fokus darauf, potenzielle Situationen zu identifizieren, welche Unregelmäßigkeiten in Bezug auf die Kündigungsrate aufzeigen. Der verwendete Datensatz weist eine Kündigungsrate von 16,3% auf, und die Visualisierungen wurden zur besseren Lesbarkeit normalisiert.

Die erste Analyse betrifft den Zusammenhang zwischen dem Geschlecht und der Zielvariable. Im vorliegenden Datensatz machen männliche Arbeitnehmer einen Anteil von 50,38% aus. Wenn wir nun das Verhältnis der Kündigungen zwischen den Geschlechtern in Abbildung 4.16 betrachten, zeigt sich eine leicht erhöhte Anzahl männlicher Kündigungen.

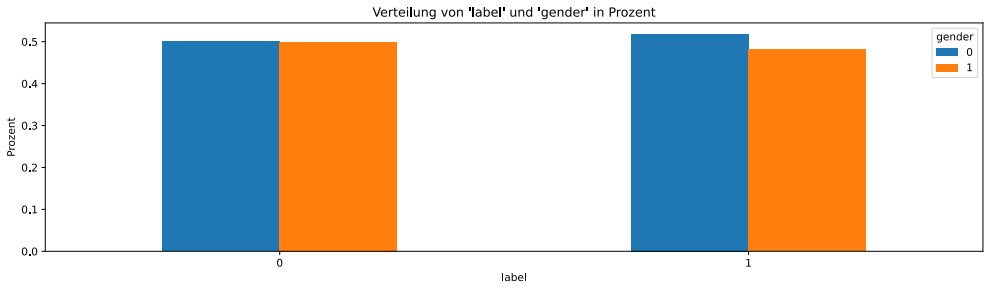


Abbildung 4.16: Kündigung im Bezug zum Geschlecht

Die Analyse des Zusammenhangs zwischen dem höchsten Schulabschluss und der Zielvariable offenbart ein vergleichbares Verhältnis zwischen Nicht-Kündigungen und Kündigungen. Auffällig ist jedoch das Verhältnis zwischen Arbeitnehmern mit Realschulabschluss (3) und Abitur (4). Hier deutet sich an, dass Arbeitnehmer mit Abitur tendenziell häufiger zu Kündigungen neigen.

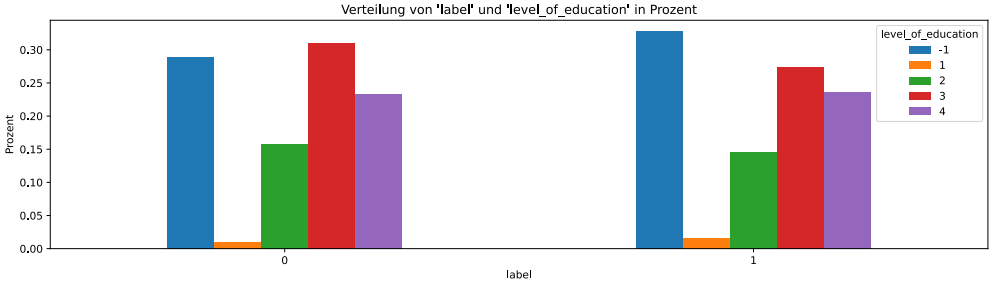


Abbildung 4.17: Kündigung im Bezug zum Schulabschluss

Der Zusammenhang zwischen dem höchsten Berufsabschluss und der Zielvariable zeigt ebenfalls ein vergleichbares Verhältnis zwischen Nicht-Kündigern und Kündigern. Es ist lediglich ein leicht höherer Kündigungsanteil von Bachelorabsolventen (4), bei einem minimal niedrigeren Kündigungsanteil von Arbeitnehmern mit Berufsausbildung (2) zu erkennen.

Folgend eine Betrachtung des Zusammenhangs zwischen Arbeitnehmer-Region und Zielvariable. In Abbildung 4.19 ist lediglich ein leichter Unterschied in der Region

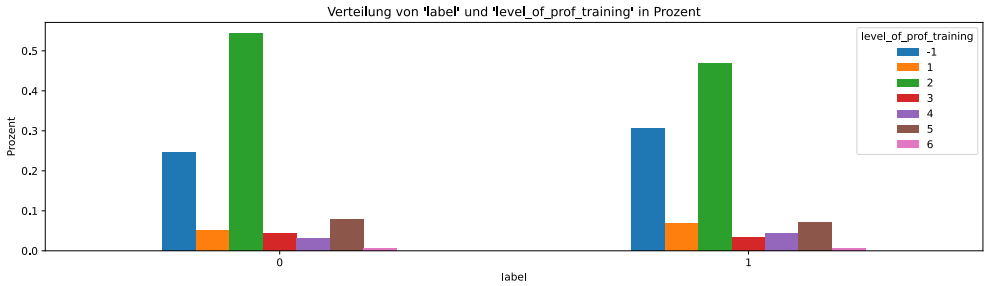


Abbildung 4.18: Kündigung im Bezug zur Berufsausbildung

„North“ zu erkennen. Hier sollte allerdings kein Zusammenhang zwischen dieser und einer Kündigung bestehen.

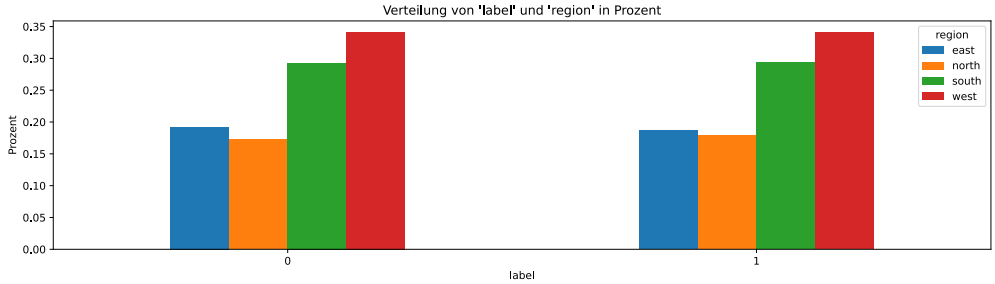


Abbildung 4.19: Kündigung im Bezug zur Region

Bei dem Vergleich des Zusammenhangs zwischen den Branchensektoren und der Zielvariable fallen insbesondere die Sektoren „Verkehr und Lagerei“ (H), „Gastgewerbe“ (I), „Erbringung von sonstigen wirtschaftlichen Dienstleistungen“ (N) und „Kunst, Unterhaltung und Erholung“ (R) auf. In diesen Sektoren ist ein vergleichsweise höherer Anteil von Kündigungen zu verzeichnen. Im Gegensatz dazu zeigen die restlichen Sektoren in diesem Datensatz keine signifikanten Auffälligkeiten, da ihre prozentualen Kündigungsanteile relativ konstant bleiben. Um Betriebsgeheimnisse zu schützen, wurde auf die Darstellung dieser Verteilung verzichtet.

Zum Ende folgt eine visuelle Darstellung des Zusammenhangs zwischen dem Bruttojahresgehalt und der Kündigungsrate in Abbildung 4.20. Deutlich ist ein klarer Trend erkennbar: Arbeitnehmer mit einem Bruttojahresgehalt unter 40.000 Euro neigen eher zur Kündigung. Besonders auffällig ist ein spürbarer Anstieg der Kündigungsrate unter einem Schwellenwert von 30.000 Euro. Es ist jedoch wichtig zu berücksichtigen, dass geringfügig Beschäftigte, im Gegensatz zu Teilzeitkräften, in diese Analyse nicht einbezogen wurden.

Die prozentualen Anteile an Kündigungen im Gehaltsbereich von 40.000 Euro bis 150.000 Euro bleiben stabil und weisen keine signifikanten Abweichungen auf. Über einem Schwellwert von etwa 150.000 Euro beginnt eine erhöhte Fluktuation. Jedoch ist diese Erscheinung auf die geringere Anzahl von Arbeitnehmern in dieser Einkommensgruppe zurückzuführen, wodurch kein direkter Zusammenhang zur Kündigungsrate nachweisbar ist.

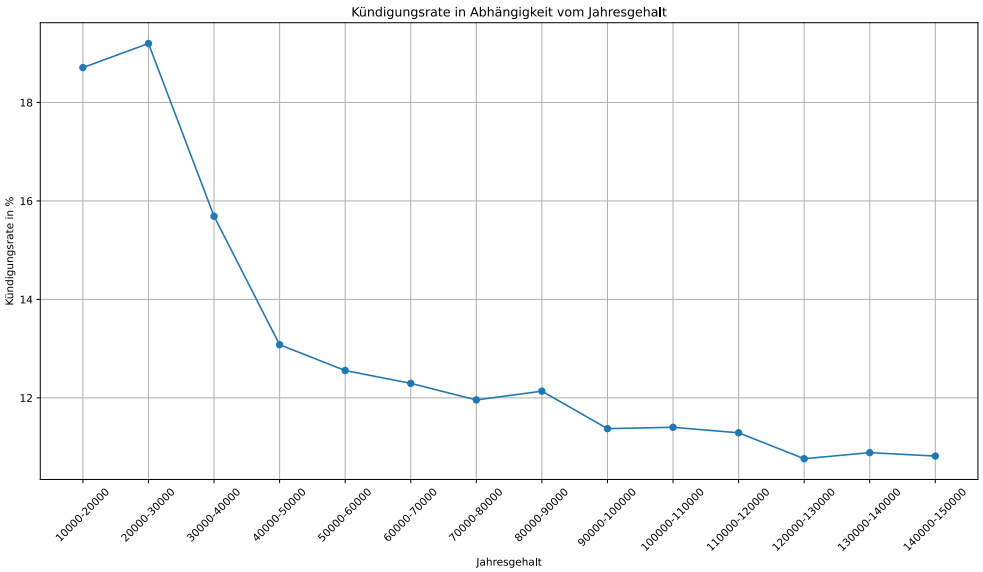


Abbildung 4.20: Kündigungsanteile im Bezug zum Bruttojahresgehalt

Abschließend präsentierte dieses Kapitel eine umfassende Datenexploration, die einen tiefen Einblick in die Struktur und Verteilung der Merkmale im Datensatz ermöglichte. Wir haben die verschiedenen Merkmale univariat und bivariat analysiert, um mögliche Zusammenhänge mit der Zielvariable, zu identifizieren. Es wurden verschiedene visuelle Darstellungen genutzt, um Muster, Trends und potenzielle Anomalien aufzuzeigen.

Es ist jedoch wichtig zu betonen, dass die vorliegenden Daten in diesem Kontext eine sehr spezifische Stichprobe darstellen. Dieser Datensatz wurde gemäß den zu Beginn des Kapitels genannten Kriterien ausgewählt und bereits bereinigt. Darüber hinaus überwiegt die Anzahl der kleinen und mittleren Unternehmen. Daher sollte klargestellt werden, dass dieser Datensatz nicht als Repräsentativ für die gesamte deutsche Bevölkerung angesehen werden kann, sondern vielmehr in dieser spezifischen Form ausschließlich für den Rahmen dieser Masterarbeit verwendet wird.

Durch diese detaillierte Untersuchung wurde ein grundlegendes Verständnis über die Bedeutung der einzelnen Merkmale und deren Einfluss auf die Kündigungsrate gewonnen. Es wird gezeigt, wie Faktoren wie Geschlecht, Alter, Berufsausbildung, Gehalt, Unternehmensgröße und Branchenzugehörigkeit in Verbindung zur Kündigung stehen. Solche Erkenntnisse bieten eine Grundlage für die spätere Modellbildung und ermöglichen es, gezielt auf diese Faktoren bei der Vorhersage von Kündigungen einzugehen.

Kapitel 5

Methodische Vorgehensweise

In diesem Kapitel wird die methodische Vorgehensweise dieser Arbeit erläutert. Zunächst wird die Systemarchitektur dargestellt, um ein grundlegendes Verständnis des Ablaufs zu vermitteln. Anschließend erfolgt eine Beschreibung der technischen Frameworks und Hilfswerkzeuge, die in dieser Arbeit verwendet werden. Zusätzlich wird die angewandte Darstellung der Analyse- und Auswertungsstrategien erläutert, darunter Techniken wie Konfusionsmatrizen, Feature Importances, Differenz zu einem Baseline-Modell sowie die Verteilung der Vorhersagemetriken der ML-Modelle. Zum Abschluss wird die Struktur des verwendeten neuronalen Netzwerks erläutert, insbesondere der Umgang mit kategorialen Merkmalen durch kategoriales Einbetten (categorical embedding).

Systemarchitektur

In Abbildung 5.1 ist die Systemarchitektur grafisch dargestellt, die aus mehreren aufeinander folgenden Phasen besteht:

- **Phase 1:** Identifizierung der am besten geeigneten Daten (Varianzanalyse, Korrelationsmatrix, Entfernung von Ausreißern usw.)
- **Phase 2:** Bereinigung und Filterung (Behandlung von Null- und fehlenden Werten)
- **Phase 3:** Auswahl der Merkmale.
- **Phase 4:** Entwicklung von Vorhersagemodellen (Random Forest und Neuronales Netz)
- **Phase 5:** Hyper-Parameter Optimierung
- **Phase 6:** Evaluation der Modelle auf den Testdatensatz (mit Hilfe der Konfusionsmatrix, AUC-Kurve, usw.)

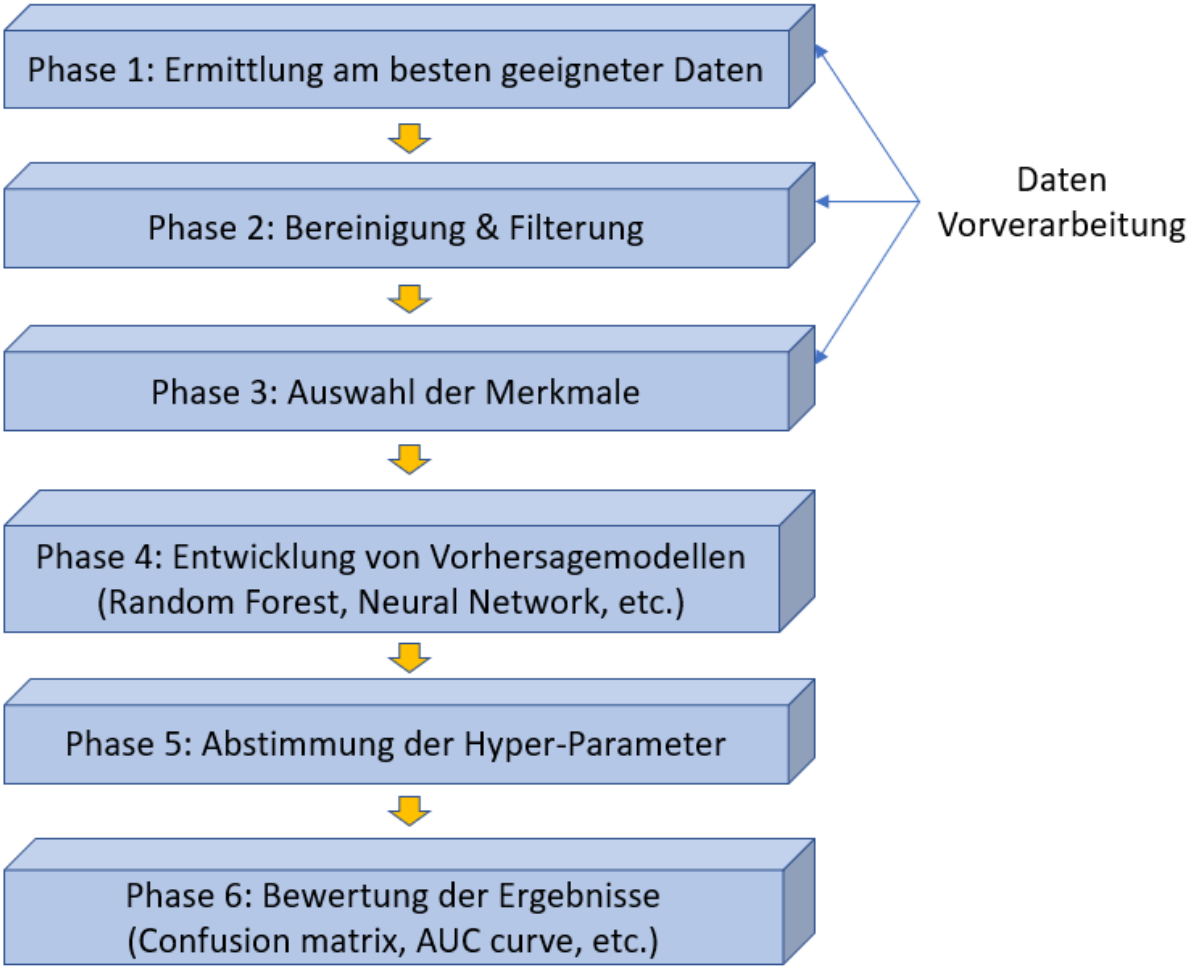


Abbildung 5.1: Mehrphasenmodell eines Frameworks für die Entwicklung eines Vorhersagemodells.

5.1 Frameworks & ML-Tools

Während des gesamten Prozesses von der Datenvorbereitung bis zur Ergebnisbewertung kam die integrierte Entwicklungsumgebung „Jupyter“ zum Einsatz, insbesondere die browserbasierte Variante Jupyter Notebook.

Zur Implementierung wurden zwei Programmiersprachen verwendet: Python und Apache Spark, speziell die pythonkompatible Version PySpark. Diese ermöglichte das Abrufen und Vorverarbeiten von Daten aus verteilten Speicherquellen. Sämtliche weiteren Schritte wurden in Python durchgeführt, wobei besonders die Software-Bibliotheken Numpy, Pandas, Scikit-learn und TensorFlow Verwendung fanden. Die Datenvorverarbeitung, Analyse und Auswertung basierten auf den tabellarisch organisierten Pandas Dataframes. Für die Modellbildung wurde vor allem auf Scikit-learn zurückgegriffen, das eine breite Auswahl an geeigneten Algorithmen (z. B. Decision Trees, Random Forest) und hilfreichen Funktionen (z. B. One-Hot-Encoding, Train-Test-Splitting) bereitstellt. Für das neuronale Netzwerk kam das Framework „TensorFlow“ zum Einsatz.

Die Visualisierungen während der explorativen Datenanalyse sowie der Datenanalyse und -auswertung der Iterationen wurden mithilfe der frei verfügbaren Software-Bibliotheken Matplotlib und Seaborn erstellt.

5.2 Darstellung der Analyse- & Auswertungsstrategien

Die Analyse und Bewertung von Klassifikationsmodellen ist essenziell, um ihre Leistungsfähigkeit im Kontext eines binären Klassifikationsproblems zu verstehen. In diesem Unterkapitel werden verschiedene Strategien zur Analyse und Auswertung dieser Modelle vorgestellt. Diese Strategien bieten tiefe Einblicke in die Fähigkeit des Modells, zwischen den beiden Klassen zu unterscheiden, und helfen, seine Stärken und Schwächen aufzuzeigen.

Während der Evaluation eines Modells ist die Konfusionsmatrix ein grundlegendes Werkzeug. Sie zeigt die Anzahl der korrekten und falschen Vorhersagen für jede Klasse. Basierend darauf werden wichtige Kennzahlen wie Genauigkeit, Präzision, Recall und F1-Score berechnet. Der F1-Score kombiniert Präzision und Recall und bietet eine ausgewogene Bewertung der Modellleistung, vor allem bei unausgewogenen Klassifikationsproblemen.

$$F1 = \frac{2 \cdot \text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}$$

Eine weitere in dieser Arbeit verwendete Metrik ist die ROC-Kurve. Sie veranschaulicht die Abhängigkeit zwischen der True Positive Rate und der False Positive Rate. Die Fläche unter der ROC-Kurve (AUC) gibt Aufschluss über die Leistungsfähigkeit des Modells.

$$\text{AUC} = \int_0^1 \text{TPR}(\text{FPR})\, d\text{FPR}$$

AUC ist die Fläche unter der Kurve (siehe Abbildung 5.2), die durch das Plotten von Sensitivität (auch als Wahre-Positiv-Rate bezeichnet) gegen die Falsch-Positiv-Rate (FPR) aufgetragen wird. TPR ist die Wahre-Positiv-Rate, die angibt, wie gut das Modell positive Fälle erkennt. FPR ist die Falsch-Positiv-Rate, die beschreibt, wie oft das Modell irrtümlicherweise negative Fälle als positiv klassifiziert. Die Integration erfolgt über den Bereich von 0 bis 1 auf der FPR-Achse. Die AUC ist ein Wert zwischen 0 und 1, wobei eine höhere AUC auf eine bessere Leistung des Modells hinweist. Ein Wert von 0,5 entspricht einem zufälligen Raten, während ein Wert von 1 darauf hinweist, dass das Modell perfekte Unterscheidungen zwischen den Klassen trifft.

Die Precision-Recall-Kurve zeigt die Beziehung zwischen Präzision und Recall. Sie ist besonders relevant, wenn die Klassenverteilung unausgewogen ist. Die Fläche unter der Precision-Recall-Kurve (AUC-PR) spiegelt die Fähigkeit des Modells wieder, positive Fälle zu identifizieren.

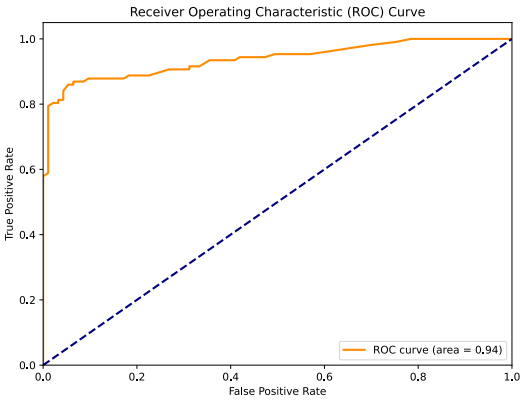


Abbildung 5.2: Beispiel einer AUC-Kurve

$$\text{AUC-PR} = \int_0^1 \text{Präzision}(\text{Wiederholungsrate})\, d\text{Wiederholungsrate}$$

Präzision ist der Anteil der korrekt positiven Vorhersagen unter allen positiven Vorhersagen. Die Wiederholungsrate ist der Anteil der tatsächlich positiven Fälle, die vom Modell erfasst werden. Die Integration erfolgt über den Bereich von 0 bis 1 auf der Wiederholungsrate-Achse. Beide Kurven bieten Einblicke in die Leistungsfähigkeit eines Klassifikationsmodells, insbesondere in Bezug auf die Balance zwischen Sensitivität und Präzision, sowie die Auswirkungen von ungleichen Klassenverteilungen auf die Vorhersagegenauigkeit.

Feature Importance

Die Analyse der Feature Importance zeigt, welche Merkmale den größten Einfluss auf die Vorhersagen haben. In Scikit-learn gibt es eine Vielzahl von Modellen, die Feature Importance berechnen können, darunter Entscheidungsbaum-basierte Modelle wie Decision Trees und Random Forests. In einem Random Forest Modell wird die Feature Importance oft auf der Grundlage von zwei Kriterien berechnet: Gini-Importance oder Mean Decrease in Impurity (MDI), und Mean Decrease in Accuracy (MDA).

- **Gini-Importance (MDI):** Dieses Maß gibt an, wie oft ein bestimmtes Feature in den Entscheidungsbäumen des Random Forests verwendet wird und wie stark es zur Verringerung der Gini-Unreinheit (eine Maßzahl für die Homogenität der Klassen in einem Knoten) beiträgt. Je größer die Verringerung der Gini-Unreinheit durch die Verwendung eines Features, desto wichtiger ist es für die Vorhersagen des Modells.
- **Mean Decrease in Accuracy (MDA):** Dieses Maß misst, wie stark die Genauigkeit des Modells abnimmt, wenn ein bestimmtes Feature aus den Daten entfernt wird. Es vergleicht die Leistung des Modells auf den ursprünglichen Daten mit seiner Leistung auf den Daten, bei denen ein bestimmtes Feature zufällig

durcheinandergeworfen wurde. Je mehr sich die Genauigkeit verringert, wenn ein bestimmtes Feature gestört wird, desto wichtiger ist dieses Feature.

Die berechnete Feature Importance wird normalerweise auf eine Skala von 0 bis 1 normalisiert, wobei 0 für keine Wichtigkeit und 1 für maximale Wichtigkeit steht. Dies ermöglicht einen direkten Vergleich der Bedeutung verschiedener Features.

Baseline-Modell

Eine weitere Evaluationsstrategie, die in dieser Arbeit angewendet wird, beinhaltet den Vergleich mit einem Baseline-Modell. Diese Vergleichsanalyse ermöglicht es, die tatsächliche Verbesserung des entwickelten Modells quantitativ zu erfassen. Im Rahmen des Baseline-Modells liegt der Fokus ausschließlich auf der Differenz des Bruttojahresgehalts, genannt „salary difference“, zwischen dem prognostizierten Gehalt und dem Marktwert. In diesem Szenario klassifiziert das Baseline-Modell den gleichen prozentualen Anteil von Arbeitnehmern wie das zu evaluierende Modell als potenzielle Kündiger. Dabei werden die Arbeitnehmer mit der niedrigsten „salary difference“ als Kündiger eingestuft. Auf Anwendungsebene haben Nutzer der PBo-Anwendung bisher nur die Marktwertprognose von Gehältern. Daher wurde diese Evaluationsstrategie gewählt, um einen optimalen Vergleich mit den bisherigen Möglichkeiten zur Vorhersage von Kündigungen zu ermöglichen.

Cross-Validation

Auch erwähnenswert ist der nutzen von Cross-Validation. Sie gewährleistet eine zuverlässige Modellbewertung und erkennt Overfitting. Darüber hinaus sorgt Hyperparameter Tuning für optimierte Modellleistung durch systematische Anpassung der Parameter. Die Analyse von Fehlern liefert Einblicke in schwierige Vorhersageszenarien. Die Visualisierung von Vorhersagen und Entscheidungsgrenzen erleichtert das Verständnis des Modellverhaltens.

Die Kombination dieser Strategien erlaubt eine umfassende Modellbewertung und stärkt das Vertrauen in die Vorhersagen. Die Modellbewertung wird dazu beitragen, gezielte Verbesserungen zu identifizieren und die Anwendbarkeit des Modells in realen Situationen zu beurteilen.

5.3 Neuronales Netz

Das in dieser Arbeit eingesetzte neuronale Netzwerk wurde eigens für die Aufgabe der Kündigungsvorhersage entwickelt. Im Vergleich zur Datenvorbereitung, wie sie im vierten Kapitel beschrieben wurde, unterscheidet sich die Vorgehensweise für dieses neuronale Netzwerk. Bei der Verwendung dieses Modells werden die kategorialen Merkmale wie „level of education“, „level of prof training“, „branch sector“ und „region“ im Gegensatz zum Random Forest nicht in One-Hot-kodierte Form umgewandelt. Sondern hier kommt eine Technik des Feature-Embeddings zum Einsatz. Zusätzlich kann nun auch die bereits erwähnte „cluster id“ in die Berechnung einfließen. Dies

war zuvor nicht möglich, da die Einbeziehung dieses Features zu einer unpraktisch hohen Anzahl von Spalten geführt hätte. Unabhängig von den hier spezifischen Daten stehen insgesamt 1301 verschiedene Cluster-IDs zur Verfügung, die eine Vielzahl von Berufen und Berufscluster repräsentieren. Um diese nun mit einfließen zu können wurde ein Ansatz eines „Multi-Input Neural Network“verwendet. Dieses Konzept ermöglicht es, verschiedene Arten von Daten (z. B. numerische, kategoriale, boolesche) in einem einzigen Modell zu verarbeiten und zu kombinieren. Die nötigen Schritte um dieses zu erstellen sind:

- **Erstellung von Modellen für verschiedene Feature-Typen:**
  - Es werden separate Modelle für numerische, kategoriale und boolesche Features mithilfe der Functional API von Keras erstellt.
  - Jedes dieser Teilmodelle beginnt mit einer Eingangsschicht, gefolgt von einer variablen Anzahl von versteckten Schichten.
  - Für das kategorische Modell müssen die kategorischen Merkmale zunächst vorberarbeitet werden. Dazu werden die verschiedenen Kategorien numerischen IDs zugeordnet. Diese IDs dienen als Eingabe für die Embedding-Schicht.
  - Für jedes kategorische Merkmal wird eine Embedding-Schicht erstellt. Diese Schicht projiziert die numerischen IDs in einen kontinuierlichen Raum niedriger Dimensionen, in dem die Beziehungen zwischen den Kategorien erlernt werden können. Die Dimension der Embedding-Schicht ist ein Hyperparameter, der experimentell angepasst werden kann.
- **Verbindung der Teilmodelle:**
  - Die Ausgangsschichten der einzelnen Teilmodelle werden durch die „Concatenate“ Schicht miteinander verbunden. Diese Schicht fügt die Ausgaben der verschiedenen Teilmodelle zusammen.
  - Die „Concatenate“ Schicht erwartet eine Liste von Tensoren, die die Ausgaben der Teilmodelle repräsentieren.
- **Erstellung des Gesamtmodells:**
  - Ein Model-Objekt wird mit Hilfe der Functional API erstellt, und die Eingangstensoren für die numerischen, kategorialen und booleschen Features werden den entsprechenden Teilmodellen übergeben.
  - Die Ausgänge der Teilmodelle werden mithilfe der „Concatenate“-Schicht verbunden.
- **Kompilierung und Training des Gesamtmodells:**
  - Das Gesamtmodell wird kompiliert, indem ein Optimierungsalgorithmus, eine Verlustfunktion und optionale Metriken festgelegt werden.
  - Anschließend wird das Gesamtmodell mit den Trainingsdaten trainiert.

Die Netzwerkarchitektur ist aufgebaut aus vier versteckten Schichten. Zur Erfassung der nicht-linearen Zusammenhänge in den Daten wurden ReLU-Aktivierungsfunktionen eingesetzt. Die Output-Schicht hingegen verwendet die Sigmoid-Aktivierungsfunktion aufgrund des Klassifizierungsproblems.

Um das Risiko von Overfitting zu mindern, wurde die Dropout-Technik in den versteckten Schichten angewandt. Als Optimierungsalgorithmus wurde der Adam Optimizer verwendet, mit einer Anfangslernrate von 0.01. Dies diente dazu, die Gewichtsanpassungen im Verlauf des Trainings zu optimieren. Zur Lösung des binären Klassifikationsproblems wurde die binäre Kreuzentropie als Verlustfunktion eingesetzt.

Das Modell durchlief initial 50 Epochen des Trainings, wobei jede Epoche eine Batch-Größe von 256 aufwies. Mit steigender Anzahl von Epochen zeigte der Validationsverlust Anzeichen von Stagnation, während der Trainingsverlust weiter abnahm. Um Overfitting zu vermeiden, wurde das Training nach 25 Epochen beendet. Durch Hyperparameter-Tuning wurden die optimalen Hyperparameter für das Modell ermittelt, um die Leistung weiter zu steigern.

In diesem Kapitel wurde ein umfassendes Verständnis der angewandten Methodik vermittelt. Dabei wurden sowohl die Systemarchitektur als auch die technischen Werkzeuge für das Training sowie die Analyse- und Auswertungsstrategien detailliert erläutert. Darüber hinaus wurde auf den Aufbau des Multi-Input Netzwerks eingegangen. Im anschließenden Kapitel erfolgt eine umfassende Diskussion der erzielten Ergebnisse sowohl aus des neuronalen Netzwerks als auch aus des Random Forest Klassifikators.

Kapitel 6

Ergebnisse und Diskussion

In diesem Kapitel werden die Resultate dieser Untersuchung präsentiert und einer eingehenden Analyse unterzogen. Dabei erfolgt ein umfassender Vergleich der Ergebnisse sowohl zwischen einem Random Forest-Klassifikator als auch einem neuronalen Netzwerk. Diese Vergleiche werden im Kontext der zuvor in Kapitel 5.2 eingeführten Metriken durchgeführt.

6.1 12 monatiger Zeitraum

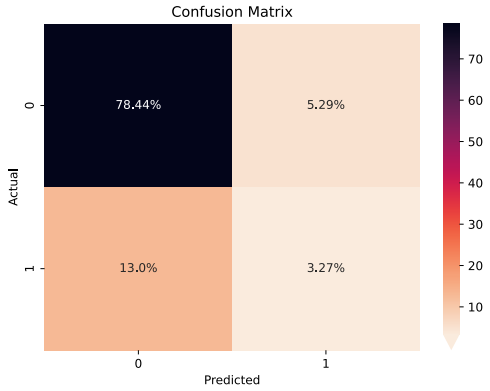
Die Ergebnisse werden für drei verschiedene Zeitspannen der Zielvariable, basierend auf dem Ansatz aus Kapitel 4.2, miteinander verglichen. Eine Vorabanalyse ergab jedoch, dass das Erlangen von guten Ergebnissen für die Zeitspanne von drei Monaten deutlich schwieriger ist. Dieses Ergebnis war hauptsächlich auf die äußerst geringe prozentuale Anzahl an Kündigungen zurückzuführen, da diese auf monatlicher Basis aggregiert werden. Dies führte zu einem Anteil von nur etwa 3% Kündigungen, was die Qualität der Ergebnisse erheblich beeinträchtigt. Aus diesem Grund wird im folgenden Abschnitten der Fokus auf die Zeitspannen von sechs bis zwölf Monaten gelegt, da diese eine bessere Datenlage und damit aussagekräftigere Ergebnisse bieten.

6.1.1 Random Forest

In Abbildung 6.1 ist eine Konfusionsmatrix dargestellt, die zusammen mit den im Kapitel 5.2 erklärten Metriken ein umfassendes Bild liefert. Die prozentuale Darstellung der Konfusionsmatrix dient der leichteren Interpretation. Zusätzlich zu den genannten Metriken enthält Tabelle 6.1a die Informationen über die Gesamtwerte der „True“ und tatsächlich „True“ Klassifikationen. Diese Analyse bezieht sich auf einen Datensatz von etwa 2 Millionen Einträgen, der im Verhältnis 80% zu 20% in Trainings- und Testdaten aufgeteilt wurde. Die Vorhersage betrifft eine aggregierte Zeitspanne von zwölf Monaten nach dem Ansatz der in Kapitel 4.2 erläutert wurde. Zusätzlich wurde ein Konfidenz-Schwellwert von 0,5 für die Bewertung festgelegt. Das bedeutet, dass das Modell eine Vorhersage nur dann als Kündigung klassifiziert wenn der Konfidenzwert

Metrik	Wert
Anteile True	0.163
Anteile True Predictions	0.086
Accuracy	0.817
Recall	0.201
Precision	0.380
F1-Score	0.79
AUC	0.69

(a) Metriken



(b) Confusion Matrix

Abbildung 6.1: Metriken des Random-Forest-Modells für einen 12 monatigen Zeitraum bei einem Konfidenz-Schwellwert von 0,5

0,5 oder höher ist. Auffallend ist, dass das Modell bei diesem Schwellwert nur 8,6% der Vorhersagen als „True“ einstuft. Hierbei ergibt sich eine Accuracy von 0,817, begleitet von einem Recall von 0,201 und einer Precision von 0,380.

In einer Problemstellung wie dieser ist eine Precision von etwa 38% als gut zu bewerten. Dies bedeutet, dass mehr als ein Drittel der Modellvorhersagen tatsächliche Kündigungen sind. Allerdings ist ein Recall von 0,201 in diesem Zusammenhang vergleichsweise niedrig, da das Modell nur 3,27% der tatsächlichen 16,3% Kündigungen erfasst. Um diesem Ungleichgewicht entgegenzuwirken und ein tieferes Verständnis für die Verteilung der Konfidenzwerte zu gewinnen, wird in der nachfolgenden Analyse einen Konfidenzschwellenwert von 0,45 untersucht.

Metrik	Wert
Anteile True	0.163
Accuracy	0.767
Anteile True Predictions	0.194
Recall	0.380
Precision	0.319
F1-Score	0.78
AUC	0.69

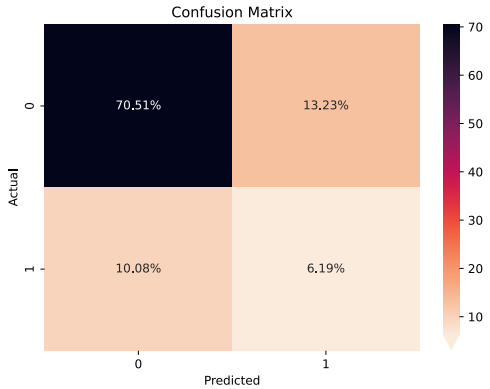


Abbildung 6.2: Metriken des Random-Forest-Modells für einen 12 monatigen Zeitraum bei einem Konfidenz-Schwellwert von 0,45

In Abbildung 6.2 sind die Ergebnisse desselben Modells bei Anwendung eines Schwellwerts von 0,45 dargestellt. Dabei beträgt der Anteil der tatsächlich positiven Vorhersagen nun 19,4%. Allerdings zeigen die Werte für die Accuracy und die Precision einen Rückgang auf 0,767 bzw. 0,319. Gleichzeitig steigt der Recall auf 0,380, wobei der F1-Score von 0,79 auf 0,78 fällt.

Diese Anpassung des Schwellwerts verdeutlicht einen Kompromiss: Es werden insgesamt mehr Vorhersagen getroffen, was sich in einem höheren Recall niederschlägt, aber gleichzeitig wird die Präzision geringer. Das bedeutet, dass eine Vorhersage nun nur noch mit einer Wahrscheinlichkeit von 31,9% eine tatsächliche Kündigung darstellt. Der F1-Score fällt von 0,79 auf 0,78. Dies zeigt, dass diese Anpassung des Schwellwertes die Balance zwischen Precision und Recall verschlechtert hat.

6.1.2 Feature Importance

Abbildung 6.3 untersucht die Bedeutung der einzelnen Merkmale für die Modellvorhersagen. Die Feature Importance vermittelt Einblicke in den Beitrag jedes Merkmals zur Vorhersagegenauigkeit des Modells. Diese Analyse ermöglicht eine genauere Bewertung der relativen Auswirkungen der verschiedenen Merkmale auf die Modellleistung. Die Darstellung der Feature Importance erfolgt anhand von quantitativen Werten, die anzeigen, wie stark jedes Merkmal zur Modellentscheidung beiträgt. Diese Werte sind normiert, sodass sie im Verhältnis zueinander stehen. Ein höherer Wert weist auf eine stärkere Beeinflussung der Modellvorhersagen durch das entsprechende Merkmal hin. Die interpretierbare Darstellung der Feature Importance unterstützt die Identifizierung von Schlüsselmerkmalen, die maßgeblich zur Modellleistung beitragen und somit einen tieferen Einblick in die Zusammenhänge zwischen den Eingabevariablen und den Zielvorhersagen ermöglichen.

Für dieses Vorhersagemodell zeigen sich besonders hohe Feature-Importance-Werte für die Merkmale „employement period“, „fluct rate“ und „age“. Insbesondere die Dauer der Beschäftigung eines Arbeitnehmers bei einem Unternehmen zeigt sich mit einem mehr als doppelt so hohen Gini-Impurity-Koeffizienten als äußerst relevant. Im Gegensatz dazu scheinen die Branchensektoren und die Berufe, in One-Hot-enkodierter Form, weniger relevant zu sein. Ebenso weisen die Felder „has short time allowance“, „supervisor“, „Region“, „wwh“, „level of education“ und „has short time allowance last year“ geringe Relevanz auf, dennoch wurden sie beibehalten. Alle gehaltsbezogenen Merkmale weisen ähnliche Gini-Impurity-Koeffizienten auf und zeigen eine gewisse Korrelation.

Ein potenziell geeignetes Verfahren an dieser Stelle ist die Principal Component Analysis (PCA) [37]. PCA ist in der Lage, die Dimensionalität eines Datensatzes zu reduzieren, indem es die Anzahl der Merkmale verringert. Gleichzeitig kann es dazu beitragen, Multikollinearität zwischen den Merkmalen zu reduzieren, was die Stabilität und Interpretierbarkeit eines Modells verbessern kann. Es ist jedoch wichtig zu beachten, dass diese Reduktion der Dimension mit einem gewissen Informationsverlust einhergehen kann. In diesem speziellen Fall wurde nach sorgfältiger Abwägung der Vor- und Nachteile beschlossen, alle Merkmale beizubehalten.

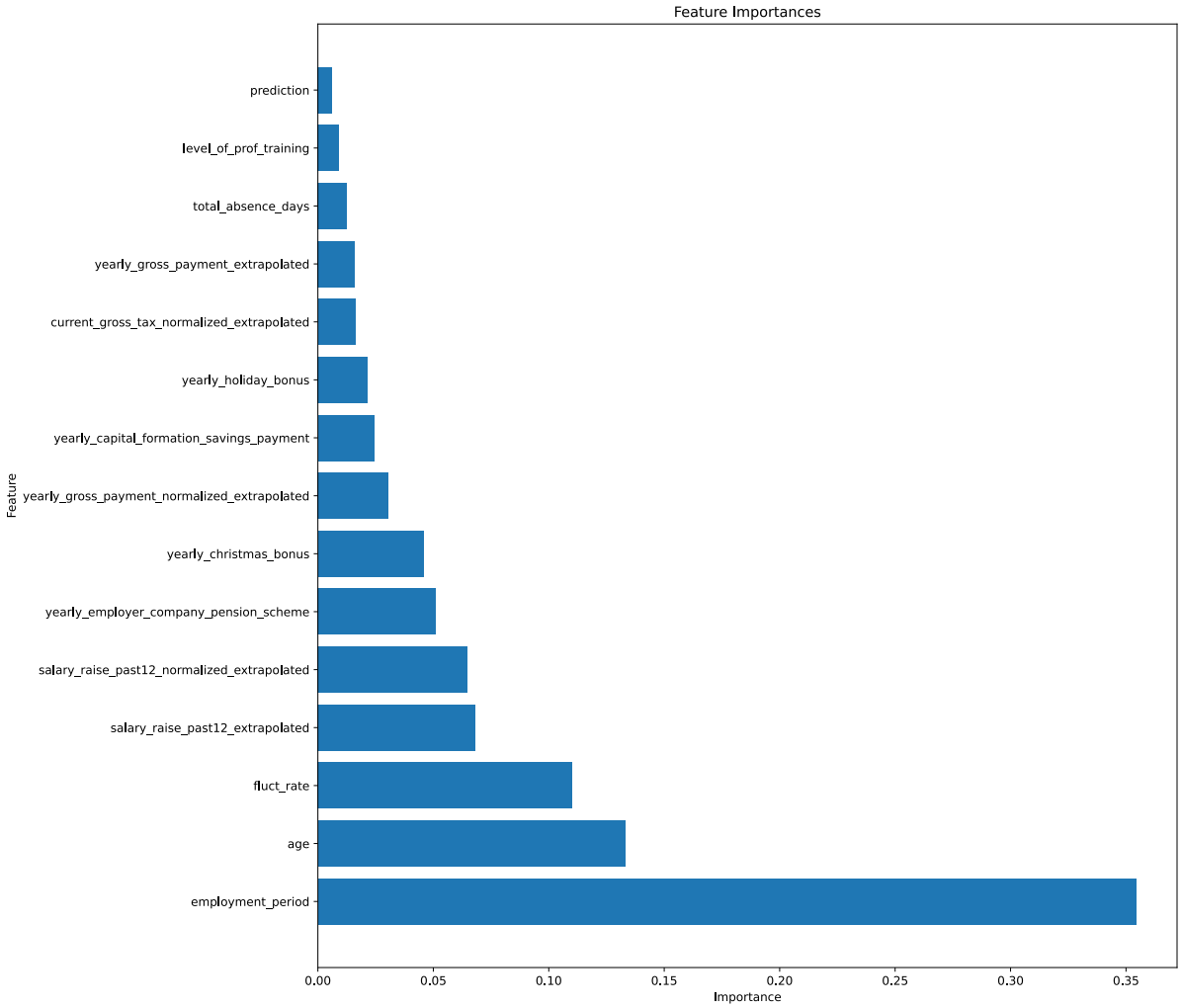


Abbildung 6.3: Die Feature Importance der Top 15 Merkmale des Random-Forest-Modells für einen 12 monatigen Zeitraum

Zusätzlich zu den bereits erwähnten Auswertungsstrategien wird in diesem Kaptel eine Ablationsstudie für die drei einflussreichsten Merkmale durchgeführt: „employement period“, „fluct rate“ und „age“. Diese Merkmale werden nacheinander entfernt, und die einzelnen Modelle werden verglichen, um die tatsächliche Relevanz dieser Merkmale besser quantifizieren zu können.

6.1.3 AUC-ROC-Kurve

In Abbildung 6.4 ist eine AUC-ROC-Kurve zusehen, welche einen AUC-Wert von 0,69 aufweist. Dieser weist eine mäßige Trennfähigkeit des Modells auf. Die Kurve zeigt den Trade-off zwischen der Rate der wahren positiven Vorhersagen und der Rate der falsch positiven Vorhersagen, während der Klassifikationsschwellenwert variiert wird. Ein höherer AUC-Wert weist auf eine bessere Modellleistung hin, da das Modell eine höhere Trennschärfe zwischen den Klassen aufweist. Die ROC-Kurve nähert sich dem optimalen Punkt oben links, was darauf hinweist, dass das Modell eine gewisse Balance zwischen Sensitivität und Spezifität erreicht hat. Dies bedeutet, dass das Modell sowohl echte Positive als auch echte Negative identifizieren kann.

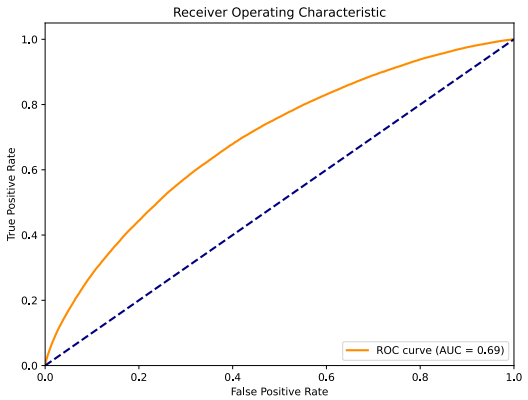


Abbildung 6.4: AUC-ROC-Kurve des Random-Forest-Modells bei einem 12 monatigem Zeitraum

6.1.4 Baseline-Vergleich

Im Zuge dieser Analyse wird ebenfalls die bereits im Abschnitt 5.2 erklärte Baseline-Bewertung berücksichtigt. In Abbildung 6.5 wird ein Baseline-Modell präsentiert, welches ausschließlich den jährlichen Gehaltsunterschied des Arbeitnehmers als Grundlage für die Vorhersage verwendet. Dieses Baseline-Modell weist einen Anteil von 8,6% positiver Vorhersagen auf, identisch wie die Ergebnisse des Modells mit einem Konfidenz-Schwellwert von 0,5. Somit ermöglicht der Vergleich dieses Baseline-Modells mit den Ergebnissen aus Abbildung 6.1 eine fundierte Einschätzung.

Hier auffällig sind die niedrigen Precision- und Recall-Werte von 0,182 bzw. 0,096. Diese Ergebnisse ergeben sich aufgrund der hohen Anzahl an „False Positives“ und „False Negatives“ in diesem Ansatz. Darüber hinaus liegt die Accuracy bei lediglich 0,782, im Vergleich zu den 0,817 in Abbildung 6.1. Der F1-Score reduziert sich von 0,79 auf 0,75 im Baseline-Ansatz.

Diese Ergebnisse verdeutlichen, dass das alleinige Verwenden des jährlichen Gehalts zur Markwertprognose als Vorhersagefaktor nicht ausreicht, um Kündigungen effektiv vorherzusagen. Daher zeigt sich die Notwendigkeit komplexerer Modelle, um eine genauere Vorhersage zu erzielen.

Metrik	Wert
Anteile True	0.163
Accuracy	0.782
Anteile True Predictions	0.086
Recall	0.096
Precision	0.182
F1-Score	0.75

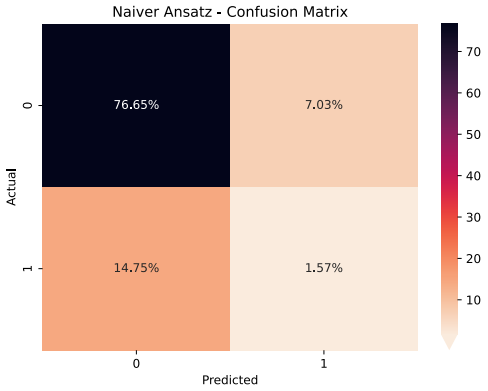


Abbildung 6.5: Metriken des Baseline-Modells als Vergleichsbasis für das Random-Forest-Modell bei einem Konfidenz-Schwellwert von 0,5

6.1.5 Neuronales Netzwerk

Aufgrund der geringen Anzahl an positiven Vorhersagen bei einem Schwellwert von 0,5 wurde die Konfusionsmatrix für diesen weggelassen. In Abbildung 6.6 ist die Konfusionsmatrix inklusive der relevanten Metriken für einen Schwellwert von 0,4 dargestellt. Anzumerken ist, dass das Random-Forest-Modell nicht direkt mit dem neuronalen Netzwerk vergleichbar ist. Vergleichbar sind nur die Ergebnisse bei denen die Anteile an positiven Vorhersagen gleich sind.

Metrik	Wert
Anteile True	0.164
Accuracy	0.835
Anteile True Predictions	0.034
Recall	0.101
Precision	0.482
F1-Score	0.79
AUC	0.71

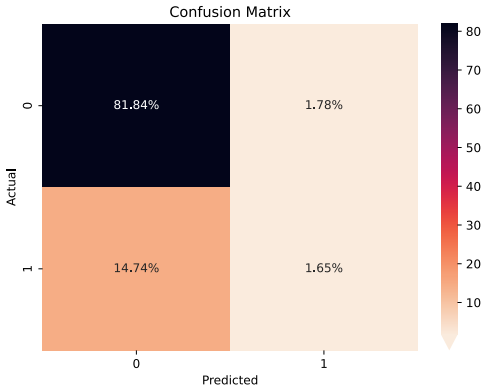


Abbildung 6.6: Metriken des neuronalen Netzwerks für einen 12 monatigen Zeitraum bei einem Konfidenz-Schwellwert von 0,4

Die erreichten Resultate bei einem Konfidenz-Schwellwert von 0,4 werden an dieser Stelle nicht unmittelbar mit dem Random-Forest-Modell verglichen. Aufgrund des niedrigem Vorhersageanteils dieses Schwellwert wird im folgedem ein Schwellwert von 0,3 betrachtet. Um eine valide Vergleichsgrundlage für den Konfidenz-Schwellwert von 0,3 zu haben, wird nun eine weitere Random-Forest-Auswertung eingeführt mit einem angepassten Schwellwert, bei welchem der Anteil an positiven Vorhersagen zwischen dem neuronalen Netzwerk bei einem Schwellwert von 0,3 und diesem gleich sind. Diese ist in Abbildung 6.7 zu sehen:

Metrik	Wert
Anteile True	0.164
Accuracy	0.808
Anteile True Predictions	0.106
Recall	0.243
Precision	0.363
F1-Score	0.79
AUC	0.69

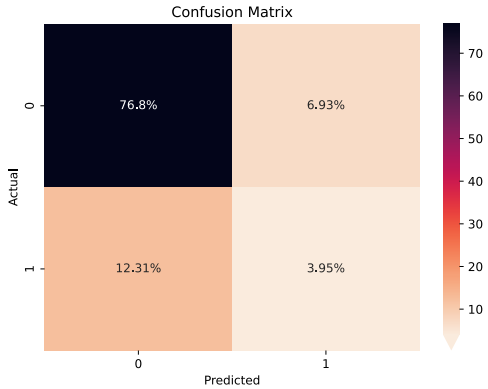


Abbildung 6.7: Metriken des Random-Forest-Modells für einen 12 monatigen Zeitraum bei einem Konfidenz-Schwellwert von 0,4856

Metrik	Wert
Anteile True	0.164
Accuracy	0.812
Anteile True Predictions	0.106
Recall	0.249
Precision	0.385
F1-Score	0.79
AUC	0.71

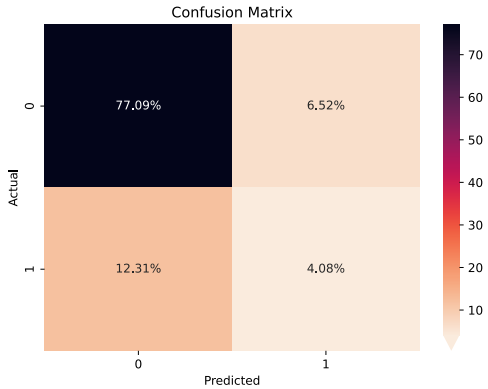


Abbildung 6.8: Metriken des neuronalen Netzwerks für einen 12 monatigen Zeitraum bei einem Konfidenz-Schwellwert von 0,3

Die Ergebnisse in Abbildung 6.8 werden nun mit denen aus Abbildung 6.7 verglichen. Es zeigt sich, dass das neuronale Netzwerk dem Random Forest überlegen ist. Die Metriken Accuracy, Recall und Precision verzeichnen alle leichte Verbesserungen im Vergleich zum Random-Forest-Modell. Der F1-Score bleibt stabil bei 0,79, jedoch steigt der AUC-Wert für das neuronale Netzwerk im Vergleich zum AUC-Wert des Random-Forest-Modells auf 0,72 von 0,69. Es ist erwähnenswert, dass die Änderungen in den Metriken Recall und Precision recht gering sind, sodass der F1-Score bei einer Rundung auf zwei Nachkommastellen keine Veränderung aufweist. Die Erhöhung des AUC-Werts bestätigt, dass das neuronale Netzwerk für den zwölfmonatigen Zeitraum dem Random-Forest-Modell überlegen ist. Es ist jedoch zu beachten, dass aufgrund der Komplexität des neuronalen Netzwerk-Modells die Konfidenzwerte niedriger ausfallen, wodurch diese Vergleichsmethode angewendet werden muss.

Grundsätzlich gestaltet sich der Vergleich von zwei Modellen, die sich in ihrer Architektur und ihren Hyperparametern grundlegend unterscheiden, als herausfordernd. In dem vorliegenden Szenario erzielen beide Modelle ähnliche Leistungen und weisen jeweils ihre eigenen Stärken und Schwächen auf. Hierunter die objektiv bessere Leistung des neuronalen Netzwerks, gegenüber der geringeren Komplexität und dadurch auch kürzeren Trainingszeit des Random Forest-Modells. Mögliche Methoden zur Auswahl der Vorhersageschwellwerte werden im Kapitel 6.4 detaillierter behandelt.

6.2 6 monatiger Zeitraum

Im folgenden Abschnitt wird die gleiche Analyse nun für einen sechsmonatigen Zeitraum der Zielvariablen-Aggregation durchgeführt.

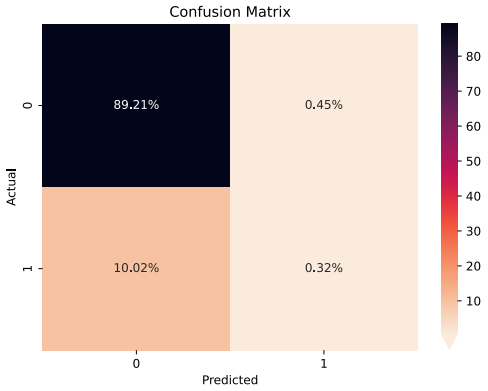
6.2.1 Random Forest

In Abbildung 6.9 ist ein Random-Forest-Modell für eine Zeitspanne von sechs Monaten zu sehen. Hierbei wurde ebenfalls ein Vorhersageschwellwert von 0,5 festgelegt. In der Grundbetrachtung zeigt sich eine höhere Accuracy von 0,897 im Vergleich zum Modell mit zwölfmonatigem Zeitraum. Es ist jedoch zu beachten, dass der Anteil an „True“-Labels bei 10,3% liegt, wodurch eine von Natur aus höhere Accuracy zu erwarten ist. Denn wenn das Modell alle Vorhersagen als 0 klassifiziert, ergibt sich bereits eine Accuracy von 0,897. Anzumerken ist dadurch, dass die Accuracy kein geeignetes Maß ist bei stark unbalancierten Datensätzen wie diesem und daher die anderen Metriken aussagekräftiger sind. Auffällig ist zudem, dass das Modell bei einem Schwellwert von 0,5 nur wenige Vorhersagen trifft. Die Precision und der Recall liegen bei 0,526 bzw. 0,052. Eine Precision von 52,6% in diesem Kontext ist respektabel, allerdings sollte beachtet werden, dass diese hohe Präzision aufgrund der geringen Anzahl positiver Vorhersagen zustande kommt. Um diesem Phänomen entgegenzuwirken, wurde ebenfalls eine Analyse des Modells mit einem Schwellwert von 0,4 durchgeführt.

In Abbildung 6.10 ist das Modell mit einem Schwellwert von 0,4 dargestellt. Hierbei zeigt sich, dass die Metriken tendenziell verschlechtert sind, während die Anzahl der

Metrik	Wert
Anteile True	0.103
Anteile True Predictions	0.010
Accuracy	0.897
Recall	0.052
Precision	0.526
F1-Score	0.86
AUC	0.72

(a) Metriken



(b) Confusion Matrix

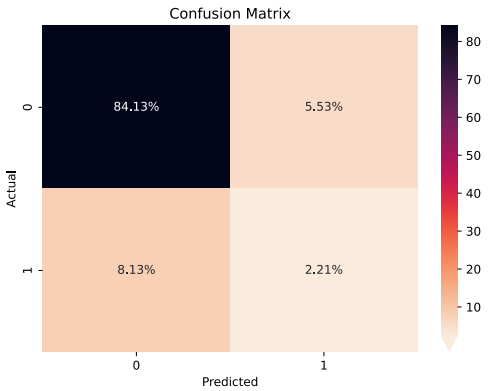
Abbildung 6.9: Metriken des Random-Forest-Modells für einen 6 monatigen Zeitraum bei einem Konfidenz-Schwellwert von 0,5

positiven Vorhersagen steigt. Die Accuracy beträgt nun 0,889 und die Precision 0,396. Dabei steigt der Recall auf 0,135. Bei der Anwendung dieser Modelle muss abgewogen werden, ob eine erhöhte Anzahl an positiven Vorhersagen den Anstieg an negativen Vorhersagen rechtfertigen kann.

Im Vergleich zum Modell mit einem 12-monatigen Zeitraum weist dieses Modell einen verbesserten F1-Score und einen höheren AUC-Wert auf. Dennoch könnte argumentiert werden, dass die Ergebnisse tendenziell schlechter sind, bedingt durch die geringere Rate an positiven Vorhersagen. Dies zeigt sich besonders im niedrigeren Recall-Wert. Dieser leicht wahrnehmbare Unterschied in der Leistung trotz nahezu identischer Hyperparameter ist im Wesentlichen auf die prozentuale Verteilung der Zielvariable zurückzuführen. Hier liegt die prozentuale Verteilung der Zielvariable bei 10,3% im Vergleich zu den 16,3% bei einem 12-monatigen Zeitraum. Im Folgenden wird versucht, diese Erkenntnisse weiter zu festigen. In Tabelle 6.4 wird dann eine abschließende Bewertung eines Vorgehensmodells präsentiert, um alle in diesem Kapitel erörterten Ergebnisse zusammenzuführen.

Metrik	Wert
Anteile True	0.103
Anteile True Predictions	0.035
Accuracy	0.889
Recall	0.135
Precision	0.396
F1-Score	0.86
AUC	0.72

(a) Metriken



(b) Confusion Matrix

Abbildung 6.10: Metriken des Random-Forest-Modells für einen 6 monatigen Zeitraum bei einem Konfidenz-Schwellwert von 0,4

6.2.2 Feature Importance

In Abbildung 6.11 sind die 15 relevantesten Merkmale des Random-Forest-Modells dargestellt. Auffällig ist, dass der Beschäftigungszeitraum nach wie vor den höchsten Importance-Score aufweist, gefolgt von der Fluktuationsrate, die ebenfalls als sehr relevant erachtet wird. Interessanterweise hat das Merkmal „Alter“ in diesem Kontext etwas an Bedeutung verloren und liegt nun auf dem fünften Platz, verglichen mit dem zweiten Platz im 12-monatigen Random-Forest-Modell. Eine weitere Auffälligkeit ist, dass nun der one-hot-encodierte Branchensektor „I“ (Gastgewerbe) zu den Top 15 Merkmalen gehört. Trotz einiger Merkmale mit geringem Importance-Score wurden auch hier alle Merkmale beibehalten, um jeglichen Informationsverlust zu vermeiden.

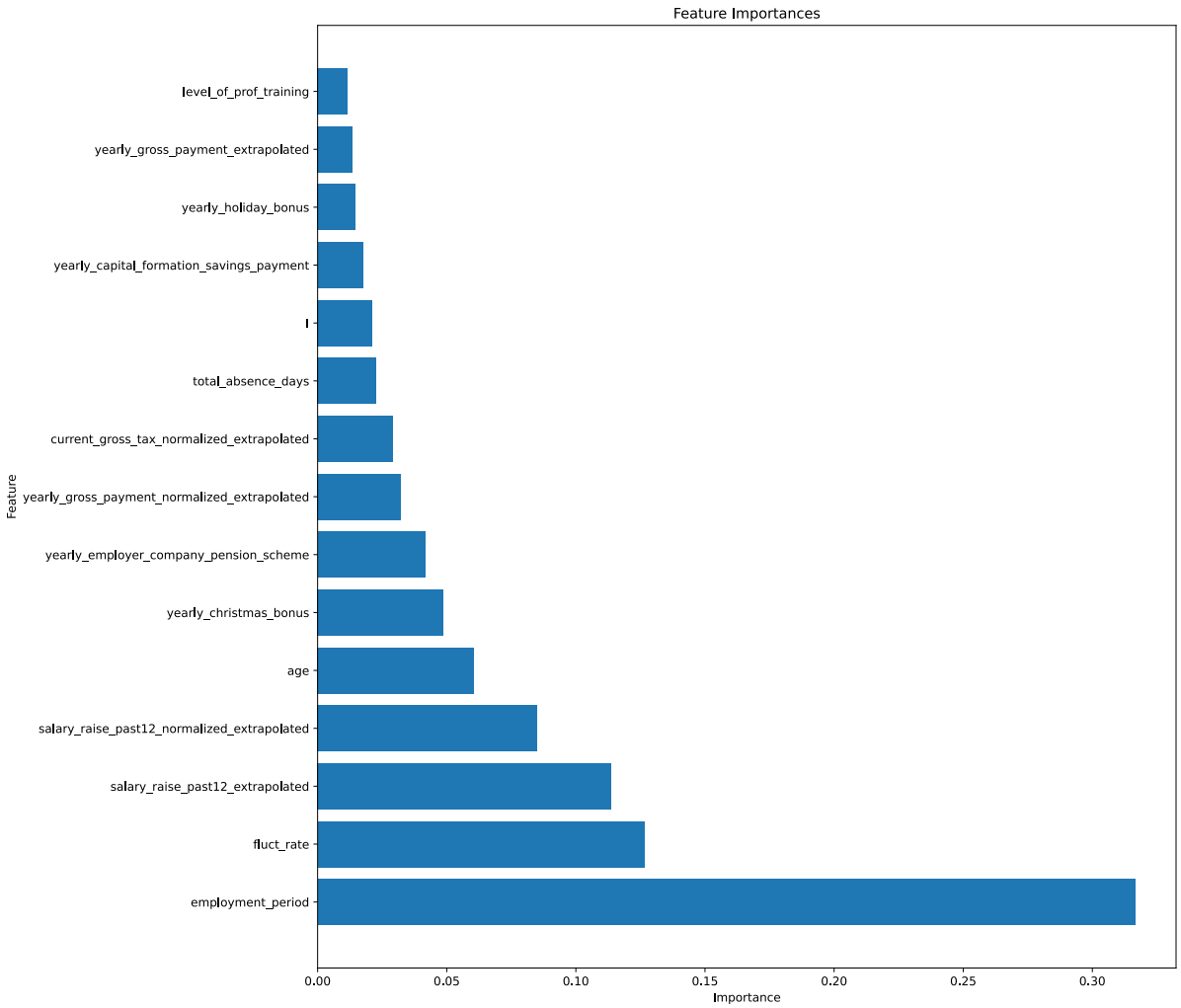


Abbildung 6.11: Die Feature Importance der Top 15 Merkmale des Random-Forest-Modells für einen 6 monatigen Zeitraum

6.2.3 AUC-ROC-Kurve

In Abbildung 6.12 ist die AUC-ROC-Kurve dieses Modells dargestellt, die im Vergleich zum Modell mit einer zwölfmonatigen Zeitspanne einen leicht verbesserten AUC-Wert

von 0,72 aufweist.

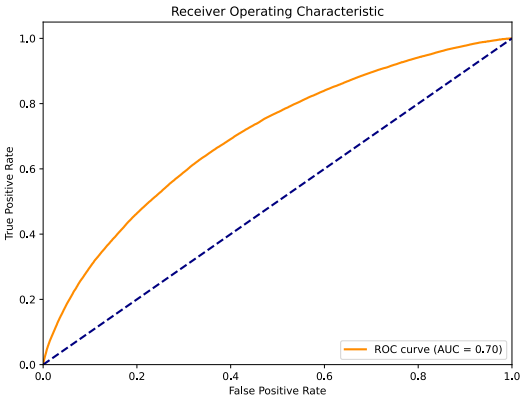


Abbildung 6.12: AUC-ROC Kurve des Random-Forest-Modells bei einem 6 monatigem Zeitraum

6.2.4 Baseline-Vergleich

Um einen fachlich fundierten Vergleich mit einem Baseline-Modell zu ermöglichen, ist es notwendig, dass sowohl das zu vergleichende Modell als auch das Baseline-Modell den gleichen Anteil an positiven Vorhersagen aufweisen. Um eine generell höhere Anzahl von positiven Vorhersagen zu erzielen und somit eine bessere Vergleichsgrundlage zu schaffen, wurde das Vorhersagemodell mit einem Schwellwert von 0,4 ausgewählt. Die Darstellung dieses Modells ist in Abbildung 6.13 ersichtlich:

Metrik	Wert
Anteile True	0.103
Accuracy	0.837
Anteile True Predictions	0.077
Recall	0.083
Precision	0.111
F1-Score	0.826

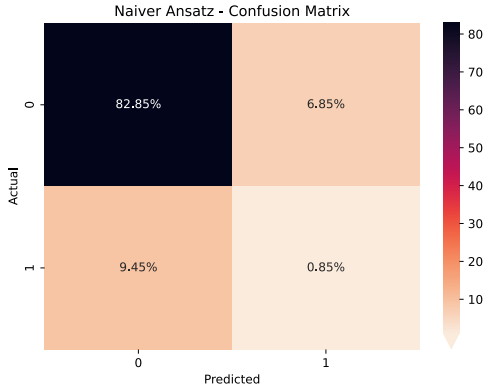


Abbildung 6.13: Metriken des Baseline-Modells als Vergleichsbasis für das Random-Forest-Modell bei einem Konfidenz-Schwellwert von 0,4

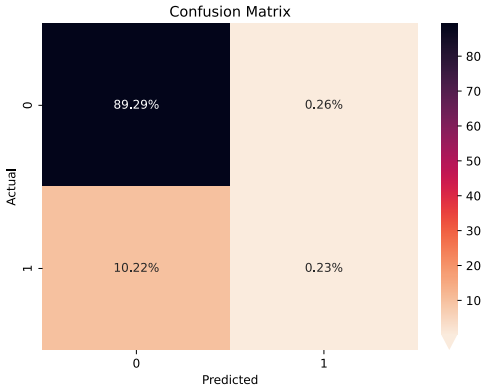
Auffällig ist hierbei, dass die Precision- und Recall-Werte mit 0,111 bzw. 0,083 niedriger ausfallen als die des Vorhersagemodells in Abbildung 6.10. Gleichzeitig zeigt die Accuracy einen Rückgang von 0,889 auf 0,837. Diese Ergebnisse verdeutlichen, dass das Vorhersagemodell unter Einbeziehung sämtlicher zuvor aufgelisteter Merkmale bessere Ergebnisse erzielt als der Ansatz, der sich ausschließlich auf das Gehalt der Arbeitnehmer konzentriert.

6.2.5 Neuronales Netzwerk

In Abbildung 6.14 ist die Konfusionsmatrix des neuronalen Netzwerks für eine Zeitspanne von 6 Monaten dargestellt. Es fällt auf, dass selbst bei einem Schwellwert von 0,4 nur ein sehr geringer Anteil von 0,5% positiven Vorhersagen zu verzeichnen ist. Im Vergleich dazu hat das Random-Forest-Modell in Abbildung 6.9 einen Anteil an positiven Vorhersagen von etwa 1%. Aufgrund dieser niedrigen Werte wird ein genauer Vergleich dieser Metriken übersprungen.

Metrik	Wert
Anteile True	0.104
Anteile True Predictions	0.005
Accuracy	0.895
Recall	0.022
Precision	0.463
F1-Score	0.85
AUC	0.71

(a) Metriken



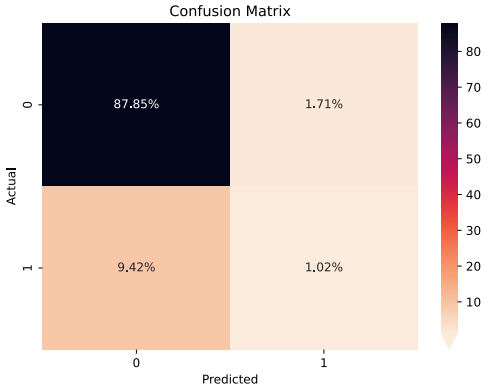
(b) Confusion Matrix

Abbildung 6.14: Metriken des neuronalen Netzwerks für einen 6 monatigen Zeitraum bei einem Konfidenz-Schwellwert von 0,4

Betrachtet man nun die Abbildung 6.15, ergibt sich bei einem Schwellwert von 0,3 ein positiver Anteil an Vorhersagen von 2,73%. Diese Resultate werden an dieser Stelle auch nicht unmittelbar mit dem Random-Forest-Modell verglichen. Die Genauigkeit beträgt 0,895, der Recall liegt bei 0,022 und die Präzision bei 0,463. Der F1-Score beträgt 0,85, während der AUC-Wert bei 0,71 liegt.

Metrik	Wert
Anteile True	0.104
Anteile True Predictions	0.027
Accuracy	0.889
Recall	0.098
Precision	0.375
F1-Score	0.86
AUC	0.71

(a) Metriken



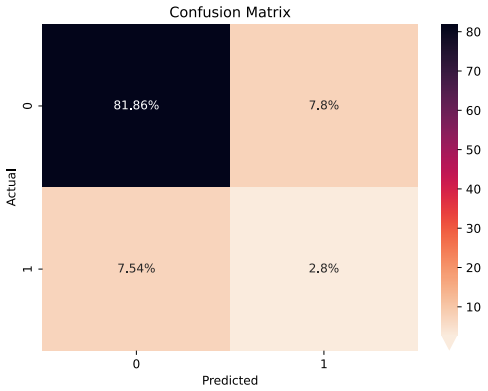
(b) Confusion Matrix

Abbildung 6.15: Metriken des neuronalen Netzwerks für einen 6 monatigen Zeitraum bei einem Konfidenz-Schwellwert von 0,3

Um eine valide Vergleichsgrundlage für den Konfidenz-Schwellwert von 0,2 zu haben, wird nun eine weitere Random-Forest-Auswertung eingeführt mit einem angepassten Schwellwert, bei welchem der Anteil an positiven Vorhersagen zwischen dem neuronalen Netzwerk und diesem gleich sind. Diese ist in Abbildung 6.16 zu sehen:

Metrik	Wert
Anteile True	0.104
Anteile True Predictions	0.106
Accuracy	0.847
Recall	0.271
Precision	0.264
F1-Score	0.85
AUC	0.72

(a) Metriken

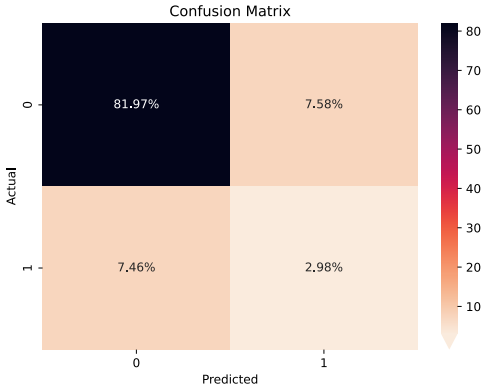


(b) Confusion Matrix

Abbildung 6.16: Metriken des Random-Forest-Modells für einen 6 monatigen Zeitraum bei einem Konfidenz-Schwellwert von 0,384

Metrik	Wert
Anteile True	0.104
Anteile True Predictions	0.106
Accuracy	0.850
Recall	0.286
Precision	0.282
F1-Score	0.85
AUC	0.71

(a) Metriken



(b) Confusion Matrix

Abbildung 6.17: Metriken des neuronalen Netzwerks für einen 6 monatigen Zeitraum bei einem Konfidenz-Schwellwert von 0,2

Bei einem Vergleich der beiden Modelle bei gleichem Anteil positiver Vorhersagen zeigt sich, dass das neuronale Netzwerk dem Random Forest überlegen ist. Die Metriken Accuracy, Recall und Precision verzeichnen alle leichte Verbesserungen im Vergleich zum Random-Forest-Modell. Der F1-Score bleibt stabil bei 0,85, jedoch fällt der AUC-Wert für das neuronale Netzwerk im Vergleich zum AUC-Wert des Random-Forest-Modells auf 0,71. Es ist erwähnenswert, dass die Änderungen in den Metriken Recall und Precision recht gering sind, sodass der F1-Score bei einer Rundung auf zwei Nachkommastellen keine Veränderung aufweist. Der Rückgang des AUC-Werts erklärt sich dadurch, dass dieser Wert über alle Konfidenzwerte berechnet wird, und der Random Forest in anderen Konfidenzwerten besser abschneidet.

In Tabelle 6.4.1 wird anschließend eine Handlungsempfehlung anhand eines Vorgehensmodells präsentiert, welches alle in diesem Kapitel aufgezeigten Ergebnisse vereinen soll. Zuvor wird die bereits erwähnte Ablationsstudie basierend auf dem Random-Forest-Modells bei einem 12-monatigem Zeitraum untersucht.

6.3 Feature-Ablationsstudie

In diesem Abschnitt wird nun genauer auf die Feature Importance aus Abbildung 6.3 eingegangen. Insbesondere werden die drei am höchsten bewerteten Merkmale der Feature Importance genauer betrachtet. Das Ziel ist es, zu verstehen, wie das Entfernen dieser Merkmale die Vorhersagefähigkeit des Modells aus Abschnitt 6.1.1 beeinflusst. Ebenso wird dieses Vorgehen nur einmal durchgeführt, da sich der Effekt des Entfernens einzelner Merkmale in Vorexperimenten als überwiegend gleich über alle Modellversuche erwies. Um Redundanz zu reduzieren wurde sich hier ausschließlich auf das Modell mit den besten Ergebnissen fokussiert.

Dieses Vorgehen, oft als „Feature-Ablationsstudie“ (Feature Ablation Study) bezeichnet, ermöglicht es, die spezifische Bedeutung jedes einzelnen Merkmals zu quantifizieren und zu verstehen, wie robust ein Modell gegenüber dem Verlust dieser Informationen ist. Die Ergebnisse dieser Studie haben weitreichende Implikationen für die Entwicklung von Modellen und den Umgang mit hochdimensionalen Datensätzen. Sie können dazu beitragen, fundiertere Entscheidungen darüber zu treffen, welche Merkmale in ein Modell aufgenommen werden sollen. Darüber hinaus kann diese Analyse unser Verständnis für die zugrunde liegenden Zusammenhänge in den Daten vertiefen und wertvolle Einblicke in die Domäne bieten, in der das Modell eingesetzt wird.

6.3.1 Random Forest – ohne die Beschäftigungsdauer

Um den Einfluss der Merkmale auf die Modellleistung genauer zu verstehen, wird zunächst das Merkmal mit dem höchsten Feature Importance Score, nämlich „Employment Period“, aus dem Modell entfernt. Die Ergebnisse dieses Experiments sind in Abbildung 6.18 in Form einer Konfusionsmatrix und entsprechender Metriken dargestellt. Die Entfernung des Merkmals „Employment Period“ führt zu einem Anstieg der Accuracy von 0,817 auf 0,829 im Vergleich zu Abbildung 6.1. Dies deutet darauf hin, dass dieses Merkmal, obwohl es einen hohen Feature Importance Score aufweist, die Gesamtgenauigkeit des Modells geringfügig Negativ beeinflusst. Ein interessanter Aspekt ist die Veränderung des Anteils an positiven Vorhersagen, der von 0,086 auf 0,046 abnimmt. Dies impliziert, dass das Modell ohne das Merkmal „Employment Period“ tendenziell weniger positive Vorhersagen trifft, jedoch mit höherer Präzision. Der Recall, der den Anteil der tatsächlich positiven Fälle abdeckt, fällt von 0,201 auf 0,115. Dies zeigt, dass das Modell nach der Entfernung des Merkmals einige positive Fälle möglicherweise nicht mehr erkennt. Die Precision hingegen steigt von 0,380 auf 0,407, was darauf hindeutet, dass die positiven Vorhersagen des Modells nach der Entfernung von „Employment Period“ tendenziell genauer sind. Der F1-Score bleibt bei 0,79.

Metrik	Wert
Anteile True	0.163
Accuracy	0.829
Anteile True Predictions	0.046
Recall	0.115
Precision	0.407
F1-Score	0.79

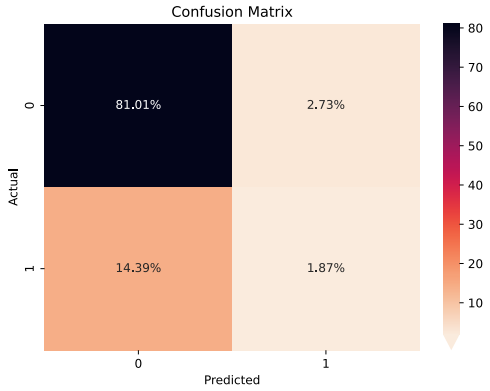


Abbildung 6.18: Metriken des Random-Forest-Modells für einen 12 monatigen Zeitraum bei einem Konfidenz-Schwellwert von 0,5 ohne die Beschäftigungsdauer

6.3.2 Random Forest – ohne die Fluktuationsrate

Die Ergebnisse des Modells ohne Berücksichtigung der Fluktuationsrate sind in Abbildung 6.19 dargestellt.

Zunächst steigt die Accuracy des Modells von 0,817 auf 0,818, was auf eine leichte Verbesserung hinweist. Der Anteil an positiven Vorhersagen nimmt von 0,086 auf 0,077 ab. Dies bedeutet, dass das Modell tendenziell weniger positive Vorhersagen macht, während die Genauigkeit leicht abnimmt. Der Recall sinkt von 0,201 auf 0,175, was darauf hindeutet, dass das Modell mit der Entfernung der Fluktuationsrate weniger in der Lage ist, tatsächliche positive Kündigungen zu identifizieren. Die Precision geht von 0,380 auf 0,371 zurück. Der F1-Score bleibt bei 0,79. Dies verdeutlicht, dass die Entfernung der Fluktuationsrate Auswirkungen auf die Modellleistung hat, insbesondere auf die Genauigkeit und Präzision. Die Ergebnisse dieser Analyse unterstreichen die Bedeutung der Fluktuationsrate als relevantes Merkmal für die Vorhersage von Kündigungen.

Metrik	Wert
Anteile True	0.163
Accuracy	0.818
Anteile True Predictions	0.077
Recall	0.175
Precision	0.371
F1-Score	0.79

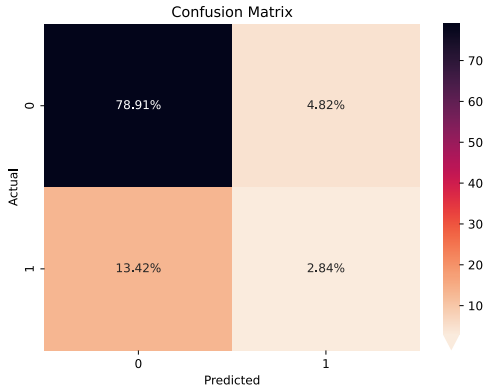


Abbildung 6.19: Metriken des Random-Forest-Modells für einen 12 monatigen Zeitraum bei einem Konfidenz-Schwellwert von 0,5 ohne die Fluktuationsrate

6.3.3 Random Forest – ohne das Alter

In Abbildung 6.20 ist die Auswertung des Modells nach Entfernung des Arbeitnehmer Alters zusehen.

Die Accuracy fällt von 0,817 auf 0,816. Der Anteil an positiven Vorhersagen verringert sich leicht auf 0,085 von 0,086„ während der Recall von 0,201 auf 0,198 sinkt. Die Precision geht von 0,380 auf 0,377 zurück, wohingegen der F1-Score konstant bei 0,79 bleibt. Diese Ergebnisse zeigen, dass das Merkmal Alter einen gewissen Einfluss auf das Modell hat, dieser jedoch geringer ist als bei der Fluktuationsrate und der Beschäftigungsdauer.

Metrik	Wert
Anteile True	0.163
Accuracy	0.816
Anteile True Predictions	0.085
Recall	0.198
Precision	0.377
F1-Score	0.79

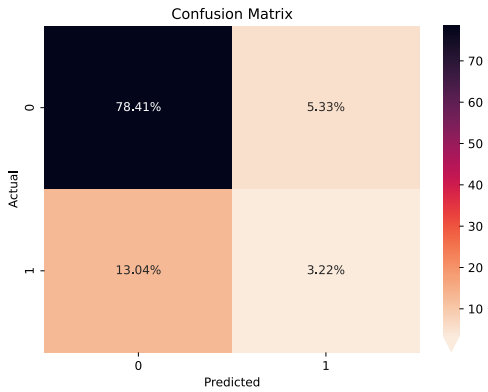


Abbildung 6.20: Metriken des Random-Forest-Modells für einen 12 monatigen Zeitraum bei einem Konfidenz-Schwellwert von 0,5 ohne das Alter

6.3.4 Fazit

Die durchgeführte Feature-Ablationsstudie hat wichtige Einblicke in die Bedeutung einzelner Merkmale für das Vorhersagemodell geliefert. Wir haben festgestellt, dass das Modell robust gegenüber der Entfernung bestimmter Merkmale ist, während es auf andere empfindlicher reagiert.

Die Entfernung des Merkmals „Beschäftigungsdauer“ führte zu einer geringfügigen Verbesserung der Accuracy, während die Entfernung der „Fluktuationsrate“ und des „Alters“ zu leichten Verschlechterungen führte. Dies verdeutlicht die unterschiedliche Relevanz dieser Merkmale für das Modell. Es unterstreicht auch die Tatsache, dass unsere Datensätze komplex sind und die Beziehung zwischen den Merkmalen und den Zielvariablen nicht immer linear ist.

Neben der genaueren Betrachtung dieser drei Merkmale wurden eine Vorabuntersuchung durchgeführt, in welcher korrelierte Merkmale, vor allem im Bezug zum Gehalt einzeln entfernt wurden. Hierunter wurden die Entfernungen des „yearly gross payment extrapolated“, „yearly gross payment normalized extrapolated“ „yearly predicted gross payment“ und „salary difference“ durchgeführt. Die einzelnen Entfernungen dieser Merkmale zeigten nur minimale Auswirkungen, die tendenziell

negativ waren. Daher wurde sich dazu entschieden, diese korrelierenden Merkmale beizubehalten.

6.4 Interpretation der Modellergenisse

Dieser Abschnitt widmet sich einer detaillierten Analyse der Ergebnisse aus den vorherigen Kapiteln. Es soll insbesondere auf die aufgetretenen Herausforderungen eingegangen und mögliche Lösungsansätze sowie Verbesserungsmöglichkeiten skizziert werden.

Basierend auf den Erkenntnissen aus den vorherigen Kapiteln lassen sich folgende Schlussfolgerungen ziehen: Ein Vorhersagemodell, das auf Lohnabrechnungsdaten beruht, ist unter den gegebenen Annahmen machbar und kann unter bestimmten Bedingungen Kündigungen vorhersagen. Die Ergebnisse zeigen auch, dass solche Vorhersagen, wenn sie auf einem komplexen Modell basieren, besser sind als die derzeit verfügbare Alternative, die allein auf Gehaltsbetrachtungen beruht. Um ein solches Modell erfolgreich umzusetzen, sind jedoch mehrere Schritte erforderlich. Zudem müssen bestimmte Bedingungen im Datensatz erfüllt sein, wie in diesem Kapitel zusammengefasst wird. Hier finden Sie auch allgemeine Hinweise zur Vorgehensweise, die je nach Datenlage variieren können. Dieses Kapitel enthält ein Modell für den Ablauf, das in Tabelle 6.4.1 verdeutlicht ist und erste Ansätze für die Verwendung von maschinellen Lernmodellen zur Vorhersage von Mitarbeiterkündigungen bietet.

Des Weiteren widmet sich dieses Kapitel den beiden Problemklassen in Klassifikationsmodellen, nämlich „False Positives“ und „False Negatives“. In einem idealen Szenario sollten die prozentualen Anteile dieser Klassen gegen Null tendieren. Bedauerlicherweise zeigen die erzielten Ergebnisse in dieser Untersuchung eine hohe Anzahl von Fehlvorhersagen auf. In den folgenden Abschnitten wird erläutert, warum diese Vorhersagen negativ ausfallen und wie sie in einem perfekten Szenario minimiert werden könnten.

Zuvor wird jedoch das Vorgehensmodell in Tabelle 6.4.1 präsentiert:

6.4.1 Vorgehensmodell für die Vorhersage von Arbeitnehmerkündigungen

Voraussetzungen:

- Repräsentative Kundendaten mit hoher Güte.
- Repräsentativer Anteil an Kündigungen (>10% ist empfehlenswert)
- Manuelle Datenaufbereitung (beispielsweise durch Ersetzung nicht vorhandener Werte im Datensatz)
- Empfehlenswert sind die 15 Merkmale mit höchstem Feature-Importance-Score.

Hinweis: Eine Unterteilung der Austritte (Kündigungen) in Kategorien und dadurch Ermöglichung konkreterer Zielvariablenbildung könnte die Ergebnisse deutlich verbessern.

Mögliches Vorgehen abhängig von Rechenleistung und Datensatz-Größe:

Trotz der langen Laufzeit und der vielfältigen Möglichkeiten sollte für jeden Klassifikator einmalig abhängig der Trainingsmenge eine Hyperparameter-Optimierung durchgeführt werden.

Unabhängig von den unten angegebenen Referenzwerte wird empfohlen, folgende Ausgangsbasen zu wählen:

- Random Forest (RF) für Vorhersagen mit hohen Konfidenz-Werten, auf Kosten der limitierten Verwendung von kategorialen Merkmalen.
- Neuronale Netzwerke (NN) für Vorhersagen mit niedrigeren Konfidenz-Werten, mit dem Vorteil kategoriale Merkmale uneingeschränkt verwenden zu können für potenziell bessere Ergebnisse.

12 monatiger Zeitraum:	**6 monatiger Zeitraum:**
• RF ist besser als NN, bei hohen Konfidenzen. • NN ist besser als RF bei niedrigen Konfidenzen. **Zusatz:** RF scheint bei dieser Konfiguration der Parameter hohe Konfidenz-Werte vorherzusagen und dadurch nur bei einem Schwellwert von 0,5 gute Ergebnisse zu liefern. NN hat hier die besten Ergebnisse bei einem Schwellwert von 0,3 und überwiegt bei Gleichstellung der positiven Vorhersagen dem RF-Modell.	• RF ist besser als NN bei hohen Konfidenzen, jedoch wenige Vorhersagen • NN ist besser als RF bei niedrigen Konfidenzen. **Zusatz:** RF liefert bei diesem Zeitraum bessere Ergebnisse, jedoch sind beide Klassifikatoren nur mäßig bei der geringeren Anzahl an positiven Zielvariablen (Kündigern). Bei Gleichstellung der positiven Vorhersagen bei niedriger Konfidenz überwiegt das NN dem RF-Modell.

Zusatz: Trainingsdauer bei beiden Klassifikationsverfahren kann abhängig von Parameterwahl lange dauern, NN tendiert allerdings zu längeren Trainingszeiten. Der Vergleich der Trainingsdauer basiert auf gleicher Datensatzgröße und gleicher Menge an Klassifikationsmerkmalen.	
Trainingsdauer/Präzision: • NN hat bessere Präzision • RF für schnellere Vorhersagen	**Trainingsdauer/Präzision:** • RF hat bessere Präzision bei Schwellwerten > 0,4, sonst hat NN bessere Präzision • RF für schnellere Vorhersagen

Tabelle 6.4.1: Vorgehensmodell für Kündigungsvorhersagen

In diesem Vorgehensmodell befindet sich eine Übersicht über die wesentlichen Voraussetzungen sowie Hinweise zur Optimierung, die auf unterschiedlichen Faktoren basieren. Darüber hinaus wurde es es in sechs- bzw. zwölfmonatige Zeiträume unterteilt, da sich die Vorhersagemodelle zwischen diesen Zeiträumen hinsichtlich ihrer Parameter und Ergebnisse unterscheiden. Dieses Modell bildet somit die Grundlage für die zukünftige Entwicklung von Vorhersagemodellen und bedarf weiterer Validierung in der Praxis. Besonders interessant ist eine Validierung anhand anderer Datensätze.

6.4.2 Interpretation der Klassifikationsfehler

Aufgrund der vorliegenden Ergebnisse und des Einsatzes eines Real-Datensatzes lassen sich einige Limitationen feststellen. Insbesondere fallen die vergleichsweise hohen Zahlen an falsch positiven und falsch negativen Vorhersagen auf. Dieses Phänomen kann weitgehend auf die Beschaffenheit des Datensatzes zurückgeführt werden. In vielen Studien im Bereich der Mitarbeiterkündigungsvorhersage werden synthetisch generierte Datensätze verwendet, die speziell für solche Problemstellungen erstellt wurden. Ein wesentliches Problem des realen Datensatzes ist das Ungleichgewicht zwischen den Klassen. Obwohl versucht wurde, diesem Ungleichgewicht entgegenzuwirken, indem Klassengewichtungen und die Aggregation der Zielvariable über mehrere Monate verwendet wurde, liegt der prozentuale Anteil an Kündigungen bei lediglich 16,3% bei einem zwölfmonatigen Zeitraum. Es ist anzunehmen, dass Vorhersagen bei Datensätzen mit natürlicherweise höheren Kündigungsraten, ohne eine solche Aggregation, besser ausfallen könnten.

Ein weiteres Limitationsmerkmal ist die Bildung der Zielvariable anhand des sogenannten „Exit"-Datums. Dieses Datum repräsentiert zwar den Zeitpunkt, an dem ein Mitarbeiter das Unternehmen verlassen hat, gibt jedoch keine Aufschlüsse darüber, aus welchen Gründen dieser Schritt unternommen wurde. Es könnten verschiedene Ursachen für eine Kündigung vorliegen, darunter:

1. **Bessere berufliche Perspektiven:** Ein besseres Jobangebot, sei es in Bezug auf Gehalt, Position oder Arbeitsbedingungen.
2. **Unzufriedenheit am Arbeitsplatz:** Gründe wie schlechtes Arbeitsumfeld, Konflikte mit Kollegen oder Vorgesetzten, Arbeitsbelastung.
3. **Umzug:** Einem Umzug aus persönlichen Gründen in eine andere Stadt oder ein anderes Land.
4. **Gesundheitliche Probleme:** Gesundheitliche Probleme, sei es physisch oder psychisch.
5. **Arbeitsplatzverlust:** Bei betrieblichen Umstrukturierungen, Fusionen oder Schließungen von Unternehmen.
6. **Fehlende Vereinbarkeit von Beruf und Familie:** Insbesondere bei jungen Eltern kann die Unfähigkeit, Beruf und Familie effektiv zu vereinbaren, zu Kündigungen führen.
7. **Karriereänderung:** Ein Mitarbeiter kann sich entscheiden, eine völlig andere berufliche Laufbahn einzuschlagen.
8. **Bessere Work-Life-Balance:** Ein Wunsch nach mehr Freizeit oder einem ausgeglicheneren Lebensstil.
9. **Burnout:** Übermäßiger Stress und Burnout-Symptome.

Wie in dieser Auflistung zu erkennen ist, gibt es eine Vielzahl von Gründen für Kündigungen. Diese neun Gründe sind sicherlich nicht vollständig und es können

noch weitere Faktoren eine Rolle spielen. Unter diesen Gründen lassen sich jedoch zwei Hauptkategorien unterscheiden: objektive Kündigungsgründe, die potenziell von KI-Modellen anhand von Lohnabrechnungsdaten vorhergesagt werden können, und subjektive Kündigungen, die nicht allein durch diese Daten erfasst werden können.

Betrachtet man diese Gründe genauer, so lässt sich annehmen, dass lediglich die Gründe 1, 2 und 8 durch unsere vorliegenden Daten repräsentiert werden können. Grund 1 könnte sich potenziell in gehaltsbezogenen Merkmalen, im Feld „Supervisor“ und in der Distanz zwischen Arbeitnehmer und Arbeitgeber widerspiegeln. Grund 2 könnte durch die Fluktuationsrate, die das Betriebsklima innerhalb eines Unternehmens repräsentiert, dargestellt werden. Ebenso relevant könnten die Abwesenheitstage im Bezug zu Unzufriedenheit am Arbeitsplatz sein. Grund 8 könnte in gewisser Weise auch durch das Distanz-Merkmal und die wöchentlichen Arbeitsstunden repräsentiert werden, obwohl beachtet werden muss, dass es bei diesem Grund noch weitere Faktoren geben kann, die nicht durch die vorhandenen Daten erfasst werden.

Die restlichen Gründe können nicht direkt anhand der vorliegenden Daten ermittelt werden. Leider erlauben es die Daten auch nicht, die Kündigungen nach rein objektiven Kündigungsgründen zu filtern, was zu einer möglichen Verbesserung der Modellgenauigkeit führen würde.

Neben potenziellen Verbesserungen in Bezug auf die Zielvariable gibt es weitere Möglichkeiten, die Leistungsfähigkeit der Vorhersagemodelle zu steigern. Ein wichtiges Merkmal, das in diesen Daten nur eingeschränkt vorhanden ist, betrifft die Abwesenheitstage eines Arbeitnehmers. Innerhalb der Daten sind lediglich die unbezahlten Abwesenheitstage erfasst. Die bezahlten Abwesenheitstage sind im Allgemeinen aussagekräftiger und könnten daher positiv zur Verbesserung der Modellgenauigkeit beitragen, da sie auch bezahlte Krankheitstage beinhalten. Bedauerlicherweise sind diese nicht innerhalb der aktuellen Datenstrukturen erfasst. Darüber hinaus gibt es weitere Einschränkungen im Bezug zur Differenzierung von Abwesenheitsgründen, wie sie in Artikel 9 der DSGVO [51] festgelegt sind. Daher ist besondere Vorsicht geboten, wenn diese Daten verarbeitet werden.

Es ist auch erwähnenswert, dass DATEV in ihren Daten bestimmte Trends in den Kundenstrukturen aufweist. Aus Gründen des Schutzes von Betriebsgeheimnissen kann hier nicht im Detail darauf eingegangen werden. Dennoch kann festgestellt werden, dass bestimmte Gruppierungen im Zusammenhang mit der Unternehmensgröße in den Daten überwiegen. Eventuell könnte eine Verbesserung der Klassifikationsqualität erzielt werden, wenn es eine gleichmäßigere Verteilung von Unternehmensgrößen und -arten gäbe.

Die verwendeten Parameter und ihre Ausprägungen, die für die Hyperparameter-Optimierungen verwendet wurden sollten weiterhin ausgebaut werden. Das verwendete neuronale Netzwerk weist viele Parameter mit unterschiedlich vielen setzbaren Ausprägungen auf. Aufgrund der zur Verfügung stehenden Kapazitäten der virtuellen Maschinen konnte dadurch nur ein Teil der Parameter und ihre Kombinationen berücksichtigt werden. Ebenfalls könnte eine Erhöhung der Komplexität des neuronalen Netzwerks zu Verbesserungen führen. Hier wurde ab einer gewissen Komplexität

bewusst gestoppt, bzw. diese wurde reduziert um Overfitting zu vermeiden. Unter bestimmten Umständen mit anderen Daten, kann eine Erhöhung zu besseren Leistungen führen. Hier jedoch besteht die Möglichkeit, dass Konfidenz-Werte bei Erhöhung der Komplexität weiter fallen, was wiederum als eine Verschlechterung gewertet werden könnte.

6.4.3 Auswahl von Konfidenz-Schwellwerten

Die Wahl des Konfidenz-Schwellwerts ist ein kritischer Schritt bei der Anwendung von Klassifikatoren wie Random Forests oder neuronalen Netzwerken. Der Standardwert für den Schwellwert beträgt oft 0,5, was bedeutet, dass eine Vorhersage mit einer Konfidenz von 0,5 oder höher als positive Vorhersage klassifiziert wird. Es gibt jedoch Situationen, in denen es sinnvoll ist, diesen Schwellenwert anzupassen, um die Leistung des Modells zu optimieren oder bestimmte Anforderungen zu erfüllen. Hier sind einige wichtige Überlegungen zur Auswahl eines geeigneten Konfidenz-Schwellwerts:

Zielsetzung berücksichtigen

Die Auswahl des Konfidenz-Schwellwerts hängt stark von den Zielen des Modells und dessen Anwendung ab. Wenn beispielsweise eine hohe Genauigkeit bei der Vorhersage von Kündigungen benötigt wird, könnte ein höher Schwellenwert gewählt werden, um sicherzustellen, dass nur sehr sichere Vorhersagen als positiv betrachtet werden. Auf der anderen Seite, wenn eine frühzeitige Erkennung von potenziellen Kündigungen priorisiert wird, könnten ein niedriger Schwellenwert gewählt werden, um empfindlicher auf positive Vorhersagen zu reagieren, auch wenn diese weniger sicher sind. Diese Prinzip wurde hier bei der Analyse der Ergebnisse in Kapitel 6 angewendet.

Kosten-Nutzen-Analyse

Das Verhältnis zwischen Kosten und Nutzen muss hier berücksichtigt werden. Das Anpassen des Schwellwerts kann dazu führen, dass mehr oder weniger falsch positive oder falsch negative Vorhersagen entstehen. Abhängig des Anwendungszwecks muss beurteilt werden, welche Art von Fehler für eine Anwendung teurer oder problematischer ist und entsprechend muss der Schwellwert dahingehend angepasst werden.

Berücksichtigung der Klassenverteilung

Die Verteilung der Zielklasse in ihren Daten kann erheblichen Einfluss auf die Auswahl des Schwellenwerts haben. Wenn die Zielklasse stark unausgewogen ist, tendiert man oft dazu, einen niedrigeren Schwellenwert zu wählen. Dies gewährleistet, dass positive Beispiele nicht übersehen werden und das Modell empfindlicher auf die Minderheit der positiven Fälle reagiert.

Zusätzlich dazu veranschaulicht Abbildung 6.21 einen Linienplot des Random-Forest-Modells für einen zwölfmonatigen Zeitraum. In dieser Grafik wird das Verhältnis

zwischen korrekten und falsch positiven Vorhersagen anhand der Precision-Metrik dargestellt. Des Weiteren sind gestrichelte Linien bei den Schwellwerten 0,4 und 0,5 eingetragen. Die durchgezogene grüne Linie repräsentiert die Gesamtzahl der positiven Vorhersagen.

Abgesehen von den im Kapitel 6 verwendeten Metriken, kann diese Grafik als Hilfestellung dienen, um einen geeigneten Schwellenwert zu ermitteln. Eine mögliche Anforderung könnte lauten, dass mindestens 10% der positiven Vorhersagen erzielt werden sollen. Dies entspräche dem Schnittpunkt des 0,1-Punkts auf der Y-Achse mit der durchgezogenen grünen Linie, welche die Gesamtzahl der positiven Vorhersagen darstellt. In diesem Fall wäre der optimale Schwellenwert somit 0,5. Ebenso könnte ein Nutzer retrospektiv festlegen, dass mindestens eine Precision von 0,4 erreicht werden soll, und dieses Prinzip rückwirkend anwenden.

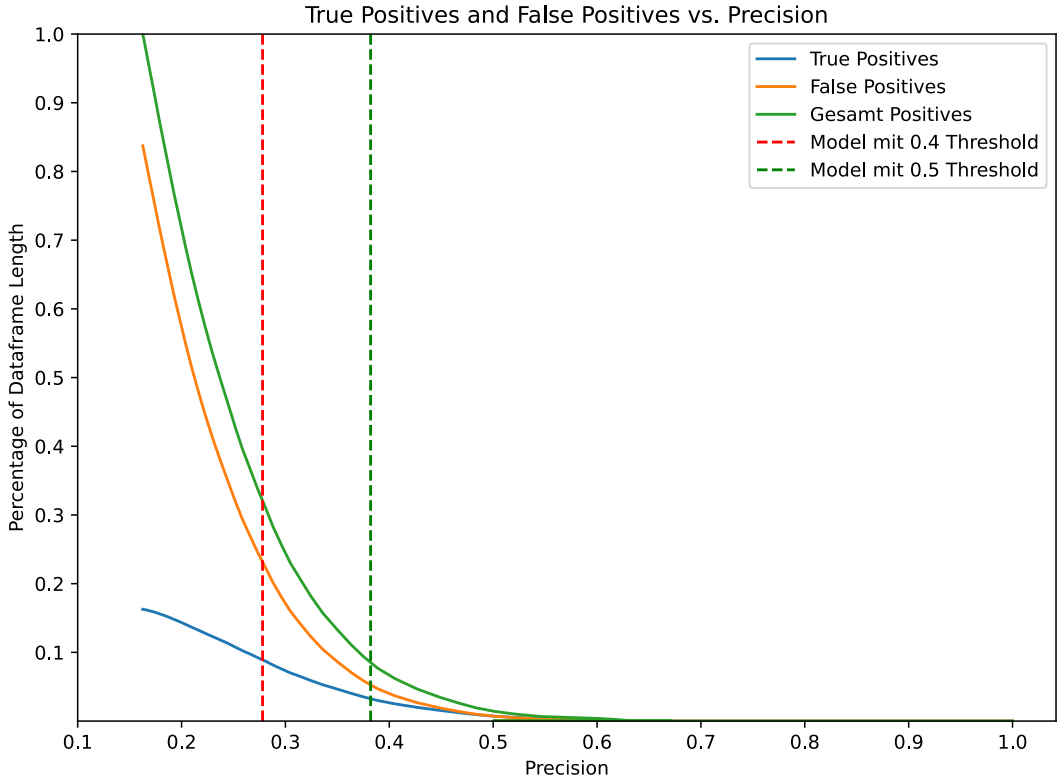


Abbildung 6.21: Verteilung der Vorhersageklassen zur Precision anhand des Random-Forest-Modells mit 12 monatigem Zielvariablen-Zeiraum.

Letztendlich hängt die Wahl des Schwellenwerts stark von dem spezifischen Anwendungsszenario des Nutzers ab. Daher wird im folgenden Kapitel 7 „Zusammenfassung & weitere Vorgehensweise“ eine Beispielanwendung skizziert. Diese verdeutlicht, wie DATEV ein Vorhersagemodell unter Verwendung verschiedener Konfidenz-Schwellwerte effektiv einsetzen könnte.

Kapitel 7

Zusammenfassung & weitere Vorgehensweise

Ausgangspunkt dieser Arbeit bildete der Mangel an Forschung zu Arbeitnehmer-Kündigungsvorhersagen basierend auf maschinellen Lernverfahren – insbesondere auf Basis von Echtdaten in großer Menge. Die vorliegende Arbeit beschäftigte sich daher mit der Frage, wie genau Kündigungen, der im Datensatz der DATEV eG enthaltenen Arbeitnehmern, mit den maschinellen Lernmethoden Random Forest und neuronalen Netzwerken vorhergesagt werden können. Dies führte zur zentralen Fragestellung, wie präzise die Kündigungen der in den DATEV eG-Datensätzen enthaltenen Arbeitnehmer mithilfe von maschinellen Lernverfahren, speziell Random Forest und neuronalen Netzwerken, vorhergesagt werden können.

Um diese Frage zu beantworten, wurde der Datensatz ausführlich erkundet und sämtliche denkbaren Merkmale identifiziert. Anschließend wurden maschinelle Lernmodelle, nämlich Random Forest und neuronale Netzwerke, miteinander verglichen. Neben den herkömmlichen Bewertungsmetriken wie Accuracy, Precision, Recall und AUC wurde ein Baseline-Modell erstellt, das zur Qualitätsanalyse beitrug. Um zukünftige Kündigungsszenarien vorherzusagen, wurden verschiedene Methoden untersucht und der Ansatz des „Learning from the Past“ angewandt. Dabei entstand eine Schlüsselfrage, wie weit in die Zukunft Kündigungen vorhergesagt werden können. Dies führte zu einer gründlichen Untersuchung verschiedener Vorhersagezeiträume, wobei sich sechs- und zwölfmonatige Zeiträume als besonders relevant erwiesen. Aufgrund der starken Ungleichgewichtung der Klassen erwies sich der dreimonatige Zeitraum als weniger sinnvoll, weshalb die Untersuchung auf sechs- und zwölfmonatige Zeiträume fokussiert wurde. Das anschließende Hyperparameter-Tuning führte zur Entwicklung der finalen Kündigungsvorhersagemodelle, die einer eingehenden Analyse und Diskussion unterzogen wurden.

Die Erkenntnisse aus den vorherigen Kapiteln legen nahe, dass ein Vorhersagemodell, das auf Lohnabrechnungsdaten basiert, unter den gegebenen Annahmen umsetzbar ist und unter bestimmten Bedingungen in der Lage sein kann, Kündigungen vorherzusagen. Die nächste wichtige Frage lautet, wie ein solches Vorhersagemodell erfolgreich

in produktive Anwendungen integriert werden kann.

Zusätzlich zu den technischen Aspekten der Implementierung ist es von entscheidender Bedeutung, die Fragen der Privacy und Sicherheit zu berücksichtigen, die mit der Verwendung von sensiblen Mitarbeiterdaten verbunden sind. In diesem Kapitel werden potenzielle Angriffsvektoren auf das Vorhersagemodell beleuchtet, um ein Bewusstsein für die damit verbundenen Risiken zu schaffen. Darüber hinaus werden mögliche Gegenmaßnahmen erörtert, die dazu beitragen können, die Sicherheit und Integrität der verwendeten Daten und des Modells zu gewährleisten.

7.1 Ein Einsatzszenario für Vorhersagemodelle: Beratungsanwendungen für Steuerberater

Grundsätzlich wäre eine Beratungsanwendung vorstellbar, die speziell auf Steuerberater zugeschnitten ist und sie bei der Beratung ihrer Mandanten in Bezug auf Fluktuation und Fachkräftemangel unterstützt. Konzeptuell könnte diese Anwendung eine Liste von Mandanten auflisten, wobei für jedes Unternehmen ein Warnsignal angezeigt wird, wenn es eine ungewöhnlich hohe Anzahl von Mitarbeitern gibt, die als kündigungsgefährdet eingestuft sind. Eine solche Anwendung könnte auch Warnungen ausgeben, die nichts mit Kündigungen zu tun haben, beispielsweise ein vergleichsweise hohes Lohnniveau, was zu hohen Kosten führt oder ein hohes Gender-Pay-Gap.

Wenn ein Steuerberater genauer auf die Liste der Arbeitnehmer eines Mandanten schaut, könnte eine Art „Ampellogik“ zur Anwendung kommen. Hierbei könnten die in Kapitel 6 behandelten Konfidenz-Schwellwerte verwendet werden. Angenommen, das Vorhersagemodell wird für jeden Arbeitnehmer in der Liste ausgeführt und gibt über eine REST-API-Schnittstelle an, ob ein Arbeitnehmer wahrscheinlich kündigt, einschließlich der Konfidenz. Anschließend könnte das Frontend visuelle Signale neben jedem Arbeitnehmer anzeigen, abhängig von der Konfidenz der Vorhersage. Zum Beispiel könnte bei sehr sicheren Vorhersagen mit einer Konfidenz von 0,6 eine leuchtend rote Ampel angezeigt werden. Bei Arbeitnehmern mit einer Konfidenz von 0,5 könnte eine orangefarbene Ampel erscheinen, und bei einer Konfidenz von 0,4 könnte eine gelbe Ampel angezeigt werden.

Im Zusammenhang mit Vorhersagemodellen für Kündigungen ist Explainable AI [52, 26] von großer Relevanz. Dies liegt daran, dass Nutzer nicht nur daran interessiert sind, Vorhersagen zu erhalten, sondern auch verstehen möchten, welche Faktoren und Merkmale zu diesen Vorhersagen geführt haben. Explainable AI ermöglicht es, die Entscheidungsfindung des Modells transparenter zu gestalten, was wiederum die Akzeptanz und das Vertrauen der Nutzer in die Vorhersagen stärkt. Dies ist besonders wichtig, wenn die Vorhersagen dazu verwendet werden sollen, gezielte Maßnahmen zur Mitarbeiterbindung zu ergreifen. Ein verständliches Modell ermöglicht es den Unternehmen, die Gründe für die Vorhersagen nachzuvollziehen und fundierte Entscheidungen zu treffen.

Aufgrund dessen könnte eine Anwendung um eine Detailansicht für jeden Arbeit-

nehmer erweitert werden, um dem Steuerberater eine umfassende Argumentationsgrundlage dafür zu bieten, warum ein Arbeitnehmer wahrscheinlich kündigen wird. Ähnlich zur Darstellung der Feature Importance könnte ein Diagramm erstellt werden, das zeigt, welche Merkmale besonders wichtig für eine spezifische Vorhersage waren. Im Gegensatz zur Feature Importance, die die Bedeutung basierend auf dem Gesamtmodell visualisiert, konzentriert sich diese Darstellung auf die Merkmale, die für die Vorhersage eines bestimmten Arbeitnehmers relevant waren. Eine geeignete Lösung hierfür könnte die SHAP (SHapley Additive exPlanations) Bibliothek [1] sein. Diese Bibliothek ermöglicht es, ein Blackbox-Maschinenlernmodell in einer für Menschen verständlichen Whitebox-Darstellung zu visualisieren.

Diese Bibliothek ist ein interpretierbares Werkzeug zur Erklärung der Vorhersagen von maschinellen Lernmodellen. Sie basiert auf der Spieltheorie und bietet eine strukturierte Möglichkeit, die Beiträge einzelner Merkmale zu den Modellvorhersagen zu quantifizieren. Die SHAP Bibliothek kann auf verschiedene Arten von Modellen angewendet werden, einschließlich linearer Modelle, Entscheidungsbäumen und neuronaler Netze. Die SHAP Bibliothek hat eine Reihe von verschiedenen Visualisierungen, unter anderem ein sogenannten Waterfall-Plot. Dieses Visualisierungstool ermöglicht es, die Einflüsse der Merkmale auf die Vorhersagen des Modells auf einen Blick zu erfassen. Es ist besonders nützlich, um die wichtigsten Merkmale zu identifizieren und ihre Auswirkungen auf positive oder negative Vorhersagen zu verstehen. Der Waterfall-Plot erstellt eine Grafik, die die Durchschnitts-Shapley-Veränderungen für jedes Merkmal anzeigt. Die Balken werden in aufsteigender Reihenfolge angeordnet, wobei die wichtigsten Merkmale oben und die weniger wichtigen unten erscheinen. Jeder Balken zeigt den Beitrag eines Merkmals zur Vorhersage und wird von einem Farbspektrum begleitet, das den Wert des Merkmals visualisiert. Dies erleichtert die Identifizierung von Mustern und Ausreißern.

Insgesamt stellt die SHAP Bibliothek ein mächtiges Werkzeug zur Interpretation von maschinellen Lernmodellen zur verfügung. Es ermöglicht es den Anwendern, die Beiträge der Merkmale zu verstehen, Muster in den Daten zu erkennen und die Modellvorhersagen auf eine erklärliche Weise zu überprüfen. Dies ist von entscheidender Bedeutung, um das Vertrauen in maschinelle Lernmodelle zu stärken und sicherzustellen, dass sie fair und verlässlich arbeiten. In Abbildung 7.1 ist ein möglicher Plot der SHAP Bibliothek zu sehen, welcher die Verständlichkeit einer Vorhersage verbessern könnte:

[1] shap.readthedocs.io. (2023): https://shap.readthedocs.io/en/latest/

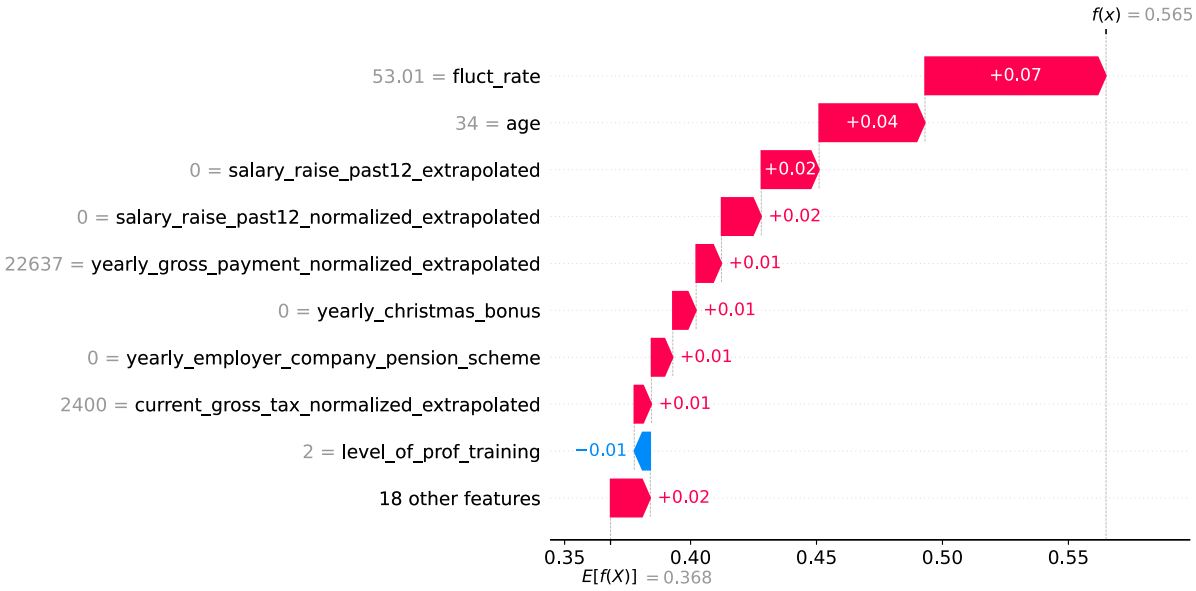


Abbildung 7.1: Eine exemplarische Ausgabe eines SHAP Waterfall-Plots für einen Random Forest bei einem zwölf monatigem Zielvariablen-Zeitraum

Das vorliegende Diagramm veranschaulicht eine positive Vorhersage, die durch das Random-Forest-Modell getroffen wurde. Es hebt hervor, dass in diesem Fall die Fluktuationsrate den größten Einfluss auf die positive Klassifizierung ausübt. Zusätzlich werden die tatsächlichen Werte in den einzelnen Datenpunkten präsentiert, um dem Benutzer eine solide Grundlage für die Interpretation zu bieten. Das Alter von 34 Jahren trägt in dieser Situation zur Neigung einer Kündigung bei. Vergleiche mit anderen Vorhersagen haben gezeigt, dass ältere Personen tendenziell weniger geneigt sind zu kündigen als jüngere Personen. Des Weiteren trägt das Fehlen einer Gehaltserhöhung in diesem Fall ebenfalls zu einem höheren Shapley-Wert bei, was wiederum auf eine Tendenz zur Kündigung hinweist. Es sollte angemerkt werden, dass die Merkmalsbezeichnungen, wie sie hier gezeigt werden, in einer tatsächlichen Anwendung möglicherweise umbenannt werden müssen. Die SHAP-Bibliothek bietet hierbei die Möglichkeit, Spaltenbezeichnungen als Parameter anzugeben.

7.2 Privacy und Sicherheit in Bezug auf Vorhersagemodelle für Kündigungen

Die Sicherstellung der Anonymisierung und Pseudonymisierung von Daten ist von höchster Bedeutung, um sicherzustellen, dass individuelle Mitarbeiter nicht identifiziert werden können. Dies ist entscheidend, um Datenschutzverletzungen zu verhindern und die gesetzlichen Anforderungen, insbesondere die Datenschutzgrundverordnung (DSGVO) der Europäischen Union, zu erfüllen.

DATEV eG, als in Deutschland ansässiges Unternehmen, unterliegt den strengen Bestimmungen der DSGVO, die die Verarbeitung personenbezogener Daten in den EU-Mitgliedstaaten reguliert und standardisiert. Die vollständige Konformitätsprüfung der entwickelten Modelle mit den komplexen Details der DSGVO kann in dieser Arbeit nicht geleistet werden. Dennoch kann eine abstrahierte Betrachtung des Themas bereits wertvolle Erkenntnisse liefern. Neben der DSGVO ist das Thema Geheimnisverrat, was im Strafgesetzbuch (StGB) geregelt ist, ein großes Problem. DATEV als Auftragsverarbeiter muss dieses Berufsgeheimnis waren.

Es ist anzumerken, dass der in dieser Arbeit verwendete Datensatz ausschließlich pseudonymisierte Personendaten enthält. Darüber hinaus hat DATEV von jedem Dateneigentümern eine Einwilligungserklärung zur Nutzung ihrer Daten zur anonymen Auswertung erhalten. Obwohl die aktuelle Produktversion von Personal Benchmark online bereits maschinelle Lernmodelle verwendet, müsste im Rahmen der Kündigungsvorhersage eine erneute Prüfung der Einwilligung durchgeführt werden.

Zusätzlich zur Datenschutzkonformität ist es von großer Bedeutung, die Daten vor möglichen Angriffsvektoren zu schützen [46, 47]. Diese könnten umfassen:

1. **Feature Tampering:** Ein Angreifer könnte versuchen, Merkmale zu ändern oder zu manipulieren, die das Modell verwendet, um Vorhersagen zu treffen. Dies könnte dazu führen, dass das Modell falsche Vorhersagen trifft.
2. **Adversarial Attacks:** In Machine-Learning-Modellen, insbesondere in neuronalen Netzen, könnten Angreifer gezielt manipulierte Eingaben erstellen, die das Modell irreführen, um falsche Vorhersagen zu erzeugen.
3. **Model Extraction:** Ein Angreifer könnte versuchen, das Modell selbst zu extrahieren, um es zu analysieren oder für böswillige Zwecke zu nutzen.
4. **Model Inversion:** Angreifer könnten versuchen, Informationen über individuelle Datensätze zu extrahieren, indem sie Vorhersagen für verschiedene Eingaben machen und die Ausgaben des Modells analysieren.
5. **Membership Inference:** Ein Angreifer versucht zu bestimmen, ob bestimmte individuelle Datenpunkte Teil des Trainingsdatensatzes eines maschinellen Lernmodells waren, indem er die Reaktionen des Modells auf Anfragen analysiert. Dies kann die Privatsphäre gefährden, indem es Rückschlüsse auf die Verwendung sensibler Daten erlaubt.

6. **Data Poisoning:** Ein Angreifer könnte versuchen, die Qualität der Trainingsdaten zu manipulieren, indem er falsche Informationen in die Datenbank einfügt. Dies könnte zu verzerrten Modellergebnissen führen.
7. **Privacy Leaks:** Das Modell könnte versehentlich sensible Informationen über Kunden oder Mitarbeiter offenbaren, wenn nicht ausreichend auf Datenschutz geachtet wird.

Adversarial & Model Inversion Attacks

Im Gegensatz zum bereits vorhandenen Modell für die Marktwertprognose bietet der Ansatz der Kündigungsvorhersage den Vorteil, dass der Benutzer bei der skizzierten Anwendung mit Warnungen oder der Ampellogik keine manuellen Eingaben innerhalb der Benutzeroberfläche vornehmen kann. Bei einer Eingabemaske, die direkt Eingaben an ein maschinelles Lernmodell weitergibt, besteht die Gefahr von Angriffen wie Adversarial Attacks und Model Inversion. Bei solchen Angriffen nutzt der Angreifer die Schnittstelle der Anwendung, sofern er keinen Zugriff auf das Modell selbst hat. Ziel ist es, durch manipulierte Eingaben falsche Ergebnisse zu erzeugen oder bei einem Model Inversion Angriff potenzielle Rückschlüsse auf vertrauliche Informationen zu ziehen.

Die Kündigungsvorhersage hingegen basiert auf einer Vielzahl von verwendeten Merkmalen und erfolgt ausschließlich auf Grundlage der vorhandenen Daten, ohne dass manuelle Eingaben in einer Anwendung erforderlich sind. Dies minimiert das Risiko solcher Angriffe erheblich. Da die Vorhersage allein auf den vorliegenden Lohnabrechnungsdaten beruht und keine direkten Nutzereingaben zulässt, sind potenzielle Angriffsvektoren wie Adversarial Attacks oder Model Inversion in diesem Szenario weniger kritisch. Dies trägt zur Sicherheit und Integrität des Vorhersagemodells bei und minimiert das Risiko von Angriffen auf die Privatsphäre der Arbeitnehmer oder die Verfälschung der Vorhersagen.

Membership Inference Attacks

Im Kontext dieser Thematik ist auch der sogenannte Membership Inference Angriff von Bedeutung. Dabei versucht ein Angreifer festzustellen, ob bestimmte individuelle Datenpunkte Teil des Trainingsdatensatzes sind, mit dem Ziel, Rückschlüsse auf darin enthaltene sensible Informationen zu ziehen. Dieses Risiko ist besonders relevant für Modelle wie die Marktwertprognose, bei denen es theoretisch möglich wäre, das Gehalt einzelner Personen oder Unternehmen zu ermitteln, was einen klaren Verstoß gegen den Datenschutz darstellen würde. Um diesem Problem entgegenzuwirken, wird beim Modell der Marktwertprognose das Verfahren des Differential Privacy eingesetzt. Differential Privacy [22] ist eine Datenschutztechnik, die entwickelt wurde, um sicherzustellen, dass individuelle Datenpunkte in einem Datensatz nicht rekonstruiert oder zurückverfolgt werden können. Sie funktioniert, indem sie absichtlich Rauschen oder Störungen zu den Daten hinzufügt, um die Identifizierung einzelner Datenpunkte zu erschweren, ohne dabei die Gesamtnützlichkeit der Daten zu beeinträchtigen.

Hingegen ist dieser Angriffstyp bei der Vorhersage von Kündigungen weniger kritisch, da hier lediglich eine binäre Klassifikation durchgeführt wird. Bei dieser Datenmenge ist es nahezu unmöglich, Rückschlüsse auf einzelne Individuen oder die in das Modell einfließenden Daten zu ziehen.

Feature Tampering & Model Extraction Attacks

Feature Tampering bezieht sich auf die Manipulation von Merkmalen oder Datenpunkten in einem Modell, um die Vorhersagen absichtlich zu beeinflussen oder zu stören. Model Extraction ist ein Angriff, bei dem ein Angreifer versucht, ein Modell zu replizieren oder wesentliche Informationen darüber zu extrahieren. Bei der DATEV eG sind diese Angriffe aufgrund umfassender Sicherheitsmaßnahmen und der spezifischen Serverarchitektur kaum bis gar nicht möglich. DATEV eG hat strenge Sicherheitsprotokolle implementiert, um die Integrität der Daten und Modelle zu schützen. Dennoch liegt die Verantwortung für die Sicherheit und den Schutz des Modells bei seinem Bereitsteller. Es ist entscheidend, dass der Anbieter des Modells sicherstellt, dass angemessene Sicherheitsvorkehrungen getroffen werden, um potenzielle Angriffe zu verhindern. Dies umfasst die Implementierung von Schutzmechanismen gegen Feature Tampering und Model Extraction sowie die regelmäßige Überprüfung und Aktualisierung dieser Schutzmaßnahmen, um sicherzustellen, dass das Modell weiterhin sicher bleibt.

Privacy Leaks

In Bezug auf mögliche Privacy Leaks wurde bereits zu Beginn dieses Kapitels eingegangen. Es ist wichtig zu betonen, dass die für das Vorhersagemodell verwendeten Daten vollständig pseudonymisiert sind und dass die Personen des Datenursprungs ausdrücklich ihre Einwilligung zur Verwendung ihrer Daten gegeben haben.

Data Poisoning Attacks

Data Poisoning stellt in diesem speziellen Szenario den kritischsten Angriffsvektor dar. Bei Data Poisoning handelt es sich um die gezielte Manipulation von Trainingsdaten, um die Leistung eines Modells absichtlich zu beeinträchtigen oder in diesem Fall, um indirekt Eingaben in das Modell zu manipulieren. In diesem Kontext könnten Angreifer versuchen, die Daten, die in das Modell einfließen, zu verfälschen, um die Vorhersagen absichtlich zu beeinflussen.

Eine Integration des Vorhersagemodells, wie in Abschnitt 7.1 erläutert, könnte in einer Anwendung für Steuerberater erfolgen. Diese Anwendungen ermöglichen jedoch auch die Erfassung neuer Mitarbeiter und könnten somit indirekt den Datenbestand beeinflussen. Das Vorhersagemodell könnte auch auf neu angelegte Mitarbeiter angewendet werden, was ähnlich wie eine manuelle Eingabe eines Nutzers ist. Hier könnten verschiedene Angriffsvektoren, wie Model Inversion und Data Poisoning, kombiniert auftreten. Dies ermöglicht indirekte Tests mit verschiedenen Eingaben und die Manipulation zukünftiger Trainingsdaten. Dies könnte potenziell dazu führen,

dass Rückschlüsse auf die Fluktuation in anderen Unternehmen gezogen werden und Datenschutzrichtlinien verletzt werden.

Um dem entgegenzuwirken, können Einschränkungen in der Anwendung implementiert werden. Zum Beispiel könnte festgelegt werden, dass Vorhersagen nur für Arbeitnehmer durchgeführt werden, die bereits mindestens einen abgerechneten Monat haben. Dadurch wird sichergestellt, dass es sich um echte Personen handelt und die Datenintegrität gewahrt bleibt.

7.3 Fazit

Zusammenfassend lässt sich festhalten, dass hinsichtlich des Datenschutzes und möglicher Angriffsszenarien im Zusammenhang mit der Integration der ML-Modelle in PBo die Datenschutzanforderungen voraussichtlich erfüllt werden können und die meisten Angriffsszenarien abgewehrt werden können. Dennoch ist die Implementierung robuster Sicherheitsmaßnahmen und kontinuierlicher Überwachungsprozesse von entscheidender Bedeutung, um die Integrität der Daten zu gewährleisten und potenzielle Angriffe zu erkennen. Die Sicherstellung von Datenschutz und Sicherheit in Vorhersagemodellen für Kündigungen erfordert eine sorgfältige Balance zwischen der Nutzung wertvoller Daten und dem Schutz der Privatsphäre der Arbeitnehmer.

Da das Themenfeld jedoch rechtliche Fragestellungen enthält, sollten die genannten offenen Punkte, wie beispielsweise die Gültigkeit von Nutzungsvereinbarungen, mit entsprechender Expertise, etwa Datenschutzbeauftragten oder Juristen, abgeklärt werden. Dies stellt sicher, dass sämtliche rechtlichen Aspekte ordnungsgemäß behandelt werden und die Anwendung im Einklang mit den geltenden Datenschutzbestimmungen steht.

Abschließend eröffnet diese Arbeit wertvolle Perspektiven für die Prognose von Mitarbeiterkündigungen und untersucht verschiedene methodische Ansätze sowie Evaluationsverfahren. Die entwickelten Modelle können potenziell Organisationen dabei unterstützen, Maßnahmen zur Mitarbeiterbindung gezielt zu planen und umzusetzen. Insgesamt leistet diese Masterarbeit einen Beitrag zur Forschung im Bereich der Mitarbeiterfluktuation und bietet wertvolle Erkenntnisse für Unternehmen, die bestrebt sind, ihre Mitarbeiterbindung zu verbessern.

7.4 Dankesagung

Abschließend möchte ich mich bei all denjenigen bedanken, die mich während der Ausarbeitung der hier vorliegenden Masterarbeit motiviert und unterstützt haben. Ohne den entsprechenden Rückhalt im Team von PBo, der DATEV eG wäre der erfolgreiche Abschluss dieser Arbeit nicht möglich gewesen. Besonderer Dank gilt hierbei meinem firmeninternen Betreuer Dr. Frank Eichinger, der mir mit viel Geduld und Hilfsbereitschaft zur Seite stand. Ebenfalls möchte ich mich bei Prof. Dr.-Ing Christoph P. Neumann und Prof. Dr. Fabian Brunner für die enge und informationsreiche Zusammenarbeit zur Ermöglichung dieser praxisorientierten Arbeit bedanken.

Literaturverzeichnis

[1] Omar Adwan, Hossam Faris, Khalid Jaradat, Osama Harfoushi, and Nazeeh Ghatasheh. Predicting customer churn in telecom industry using multilayer preceptron neural networks: Modeling and analysis. *Life Science Journal*, 11:75–81, 01 2014.

[2] Abdelrahim Kasem Ahmad, Assef Jafar, and Kadan Aljoumaa. Customer churn prediction in telecom using machine learning in big data platform. *Journal of Big Data*, 6(1):28, Mar 2019.

[3] Qasem Al-Radaideh and Eman Alnagi. Using data mining techniques to build a classification model for predicting employees performance. *International Journal of Advanced Computer Science and Applications*, 3, 02 2012.

[4] Andry Alamsyah and Nisrina Salma. A comparative study of employee churn prediction model. In *Proceedings of the International Conference on Software Technology and Computer (ICSTC)*, pages 1–4, 08 2018.

[5] David O. Alao and A. B. Adeyemo. Analyzing employee attrition using decision tree algorithms. 2013.

[6] Pradeep Asthana. A comparison of machine learning techniques for customer churn prediction. *International Journal of Pure and Applied Mathematics*, 119(10):1149–1169, 2018.

[7] Cameron Barbee, Tim Hoffmann, Christian Piffel, Tobias Schotter, Sebastian Schuscha, Philipp Stangl, Thomas Stangl, and Christoph P. Neumann. FireForceDefense: Graphisches Tower-Defense-Spiel mit Kubernetes-Deployment. Technical Report CL-2021-05, Ostbayerische Technische Hochschule Amberg-Weiden, CyberLytics-Lab an der Fakultät Elektrotechnik, Medien und Informatik, 7 2021.

[8] James Bergstra and Yoshua Bengio. Random search for hyper-parameter optimization. *Journal of Machine Learning Research*, 13:281–305, 2012.

[9] Albert Bifet, Geoff Holmes, Bernhard Pfahringer, and Richard Kirkby. New ensemble methods for data streams. *Machine Learning*, 69(2-3):345–376, 2007.

[10] Christopher M. Bishop. *Pattern Recognition and Machine Learning*. Springer, 2006.

[11] Andrew P. Bradley. The use of the area under the roc curve in the evaluation of machine learning algorithms. *Pattern Recognition*, 30(7):1145–1159, 1997.

[12] Ionut Brandusoiu, G. Toderean, and Horia Beleiu. Methods for churn prediction in the pre-paid mobile telecommunications industry. In *Proceedings of the International Conference on Communications (ICComm)*, pages 97–100, 06 2016.

[13] Dariusz Brzeziński, Jerzy Stefanowski, and Szymon Wilk. Reacting to different types of concept drift: The accuracy updated ensemble algorithm. *Information Sciences*, 277:137–155, 2014.

[14] Statistisches Bundesamt. Klassifikation der wirtschaftszweige. `https://www.destatis.de/DE/Methoden/Klassifikationen/Gueter-Wirtschaftsklassifikationen/Downloads/klassifikation-wz-2008-3100100089004-aktuell.pdf?__blob=publicationFile`.

[15] Statistisches Bundesamt. Methodenbericht - interaktiver gehaltsvergleich. `https://www.destatis.de/DE/Service/Statistik-Visualisiert/Gehaltsvergleich/Methoden/Methodenbericht.html`, 2020.

[16] Jonathan Burez and Dirk Van den Poel. Handling class imbalance in customer churn prediction. *Expert Systems with Applications*, 36:4626–4636, 06 2008.

[17] Tianqi Chen and Carlos Guestrin. Xgboost: A scalable tree boosting system. In *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*, pages 785–794, 08 2016.

[18] Chen-Fu Chien and Li-Fei Chen. Data mining to improve personnel selection and enhance human capital: A case study in high-technology industry. *Expert Systems with Applications*, 34:280–290, 01 2008.

[19] Gregory Ditzler, Manuel Roveri, Cesare Alippi, and Robi Polikar. Learning in nonstationary environments: A survey. *IEEE Computational Intelligence Magazine*, 10(4):12–25, 2015.

[20] Adriaan Dries and Mykola Pechenizkiy. A review on handling concept drift in big data. *Big Data Research*, 11:1–17, 2018.

[21] Adriaan Dries and Daan Westra. Online concept drift detection on data streams. In *Proceedings of the 2014 International Symposium on Intelligent Data Analysis (IDA)*, pages 95–106, 2014.

[22] Cynthia Dwork. Differential privacy. In Michele Bugliesi, Bart Preneel, Vladimiro Sassone, and Ingo Wegener, editors, *Automata, Languages and Programming. ICALP 2006*, volume 4052 of *Lecture Notes in Computer Science*. Springer, 2006.

[23] F. Eichinger and M. Mayer. Predicting salaries with random-forest regression. In B. Alyoubi, C. E. Ben Ncir, I. Alharbi, and A. Jarboui, editors, *Machine Learning and Data Analytics for Solving Business Problems. Unsupervised and Semi-Supervised Learning*. Springer, Cham, 2022.

[24] João Gama and et al. A survey on concept drift adaptation. *ACM Computing Surveys (CSUR)*, 46(4):1–37, 2014.

[25] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep Learning*. MIT Press, 2016.

[26] David Gunning, Mark Stefik, Jaesik Choi, Timothy Miller, Simone Stumpf, and Guang-Zhong Yang. Xai—explainable artificial intelligence. *Science Robotics*, 4:eaay7120, 12 2019.

[27] Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer, 2009.

[28] Johannes Horst, Manuel Zimmermann, Patrick Sabau, Saniye Ogul, Stefan Ries, Tobias Schotter, and Christoph P. Neumann. OPCUA-Netzwerk: Angular- und FastAPI-basierte Entwicklung eines OPC-UA Sensor-Netzwerks für den Heimbereich. Technical Report CL-2023-01, Ostbayerische Technische Hochschule Amberg-Weiden, CyberLytics-Lab an der Fakultät Elektrotechnik, Medien und Informatik, 3 2023.

[29] Bingquan Huang, Mohand Tahar Kechadi, and Brian Buckley. Customer churn prediction in telecommunications. *Expert Systems with Applications*, 39(1):1414–1425, 2012.

[30] Yiqing Huang, Fangzhou Zhu, Mingxuan Yuan, Ke Deng, Yanhua Li, Bing Ni, Wenyuan Dai, Qiang Yang, and Jia Zeng. Telco churn prediction with big data. In *SIGMOD Conference 2015*, pages 607–618, 2015.

[31] Ron Kohavi. A study of cross-validation and bootstrap for accuracy estimation and model selection. *International Joint Conference on Artificial Intelligence*, 14(2):1137–1145, 1995.

[32] Bartosz Krawczyk, Michal Wozniak, and Gerald Schaefer. Cost-sensitive learning with deep neural networks for imbalanced classification. *IEEE Transactions on Neural Networks and Learning Systems*, 28(8):1820–1837, 2017.

[33] Ludmila I. Kuncheva. *Combining Pattern Classifiers: Methods and Algorithms*. John Wiley & Sons, 2004.

[34] Praveen Lalwani, Manas Mishra, Jasroop Chadha, and Pratyush Sethi. Customer churn prediction system: a machine learning approach. *Computing*, 104:1–24, 02 2022.

[35] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *Nature*, 521(7553):436–444, 2015.

[36] Jia et al. Lu. Learning under concept drift: A review. *IEEE Transactions on Knowledge and Data Engineering*, 31:2346–2363, 2019.

[37] Andrzej Maćkiewicz and Waldemar Ratajczak. Principal components analysis (pca). *Computers & Geosciences*, 19:303–342, 1993.

[38] John McCarthy, Marvin L. Minsky, Nathaniel Rochester, and Claude E. Shannon. A proposal for the dartmouth summer research project on artificial intelligence. `https://www.aaai.org/ojs/index.php/aimagazine/article/view/1904`, 1955.

[39] Tom M. Mitchell. *Machine Learning*. McGraw Hill, 1997.

[40] Imran Muhammad and Zhijun Yan. Supervised machine learning approaches: A survey. *ICTACT Journal on Soft Computing*, 5(3), 2015.

[41] Nils J. Nilsson. *Artificial Intelligence: A New Synthesis*. Morgan Kaufmann, 2014.

[42] Sinno Jialin Pan and Qiang Yang. A survey on transfer learning. *IEEE Transactions on Knowledge and Data Engineering*, 22(10):1345–1359, 2010.

[43] Pavansubhash. Ibm hr analytics employee attrition & performance. `https://www.kaggle.com/datasets/pavansubhasht/ibm-hr-analytics-attrition-dataset`.

[44] Punnoose Rohit and Pankaj Ajit. Prediction of employee turnover in organizations using machine learning algorithms. *International Journal of Advanced Research in Artificial Intelligence*, 5, 10 2016.

[45] Stuart J. Russell and Peter Norvig. *Artificial Intelligence: A Modern Approach*. Pearson Education, 2016.

[46] Patrick Sabau. Analyse und Bewertung von Methoden zur Sicherung der Vertraulichkeit in Neuronalen Netzen. Master's thesis, Ostbayerische Technische Hochschule Amberg-Weiden, September 2023.

[47] Patrick Sabau and Christoph P. Neumann. Analyse von methoden zur sicherung der vertraulichkeit in neuronalen netzen. Technical Report 2024, Ostbayerische Technische Hochschule Amberg-Weiden, March 2014. Accepted for publication.

[48] Randall Sexton, Shannon McMurtrey, Joanna Michalopoulos, and Angela Smith. Employee turnover: A neural network solution. *Computers & Operations Research*, 32:2635–2651, 10 2005.

[49] Amir Mohammad Esmaieeli Sikaroudi, Rouzbeh Ghousi, and Ali Esmaieeli Sikaroudi. A data mining approach to employee turnover prediction (case study: Arak automotive parts manufacturing). *Journal of Industrial and Systems Engineering*, 8:106–121, 2015.

[50] Dominik Smrekar, Johannes Horst, Patrick Sabau, Saniye Ogul, Tobias Schotter, and Christoph P. Neumann. OTH-Wiki: Ein Angular- und FastAPI-basiertes Wiki für Studierende. Technical Report CL-2022-04, Ostbayerische Technische Hochschule Amberg-Weiden, CyberLytics-Lab an der Fakultät Elektrotechnik, Medien und Informatik, 7 2022.

[51] Europäische Union. Art. 9. dsgvo. `https://dsgvo-gesetz.de/art-9-dsgvo/`, 2023.

[52] Feiyu Xu, Hans Uszkoreit, Yajing Du, Wenhao Fan, Dongyan Zhao, and Jun Zhu. Explainable ai: A brief survey on history, research areas, approaches and challenges. In Jie Tang, Min-Yen Kan, Dongyan Zhao, Shuo Li, and Hui Zan, editors, *Natural Language Processing and Chinese Computing. NLPCC 2019*, volume 11839 of *Lecture Notes in Computer Science*. Springer, 2019.

[53] İbrahim Yiğit and Hamed Shourabizadeh. An approach for predicting employee churn by using data mining. In *Proceedings of the IEEE International Conference on Data Analytics and Processing (IDAP)*, 09 2017.

[54] M. Zacher. Mit modernen business-anwendungen veränderungen meistern. `https://www.datev-magazin.de/archiv/mit-modernen-business-anwendungen-veraenderungen-meistern-101769`, 2023.

[55] Yue Zhao, Maciej K. Hryniewicki, Francesca Cheng, Boyang Fu, and Xiaoyu Zhu. Employee turnover prediction with machine learning: A reliable approach. In *Intelligent Systems with Applications*, 2018.

Abbildungsverzeichnis

Tabellenverzeichnis